September 1982
*Also numbered AIM-347*

Report No. STAN-CS-82-930

2

# Depth from Edge and Intensity Based Stereo

by

Henry Harlyn Baker

Department of Computer Science

Stanford University
Stanford, CA 94305

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

| REPORT DOCUMENTATION PAGE | | READ INSTRUCTIONS BEFORE COMPLETING FORM |
|---|---|---|
| 1. REPORT NUMBER<br>STAN-CS-82-930, also AIM-347 | 2. GOVT ACCESSION NO.<br>[illegible] | 3. RECIPIENT'S CATALOG NUMBER |
| 4. TITLE (and Subtitle)<br>Depth from Edge and Intensity Based Stereo | | 5. TYPE OF REPORT & PERIOD COVERED<br>Thesis August 1978 throug September 1981 |
| | | 6. PERFORMING ORG. REPORT NUMBER |
| 7. AUTHOR(s)<br>Henry Harlyn Baker | | 8. CONTRACT OR GRANT NUMBER(s)<br>MDA 903 C 0102 |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS<br>Department of Computer Science,<br>Stanford University, Stanford, CA 94305 | | 10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS<br>Arpa Order 2494 |
| 11. CONTROLLING OFFICE NAME AND ADDRESS<br>Tesident representative, Office of Naval Research, 165 Durand, Stanford University, Stanford CA | | 12. REPORT DATE<br>September 1982 |
| | | 13. NUMBER OF PAGES<br>98 pages |
| 14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) | | 15. SECURITY CLASS. (of this report) |
| | | 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE |

16. DISTRIBUTION STATEMENT (of this Report)

This document is approved for public release and sale; Distribution is unlimited. This document may be reproduced for any purpose of the United States Government.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Stereo vision, Depth map, edge correspondence, Cross-correspondence, Dynamic programming, Viterbi correlation, parallel implementations.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This document describes a system for obtaining depth maps from a stereo pair of images. Its analysis is based on techniques for subsequent intensity correspondence. The process functions on a line-by-line basis, and produces a full depth map for the viewed scene. The analysis of two example scenes is presented.

# ACKNOWLEDGMENTS

I have benefitted a great deal over the last five years of my graduate work from proximity to[1] many bright young[2] minds. At Stanford, Peter Blicher,[3] Allan Miller, Michael Lowry, Rod Brooks, Peter Hochschild, Dave Marimont, David Lowe, David Arnold and Sid Liebes have given me, at various times, help, encouragement, and plenty of distraction. Of course most of our work and all of our theses will owe a big debt to Hans Moravec for the graphics environment and picture utilities he provided us. At Illinois, Chuck Jacobus, Michael Selander, and Rod Fletcher were the principals in an exciting, though short-lived, vision group — working and discussing with them helped make Urbana enjoyable.

Dave Waltz has served me wonderfully as an advisor, supporter, and bureaucratic untangler.

Most of my thanks must go to Tom Binford, who threw me a lifeline at the end of my second year in Illinois, when funding problems at the Coordinated Science Laboratory made my continuing in vision there impossible. His intellect and scientific determination make him stand out in the field, and pose the Stanford effort in robotics and machine vision as the most coherent in the country. I have gained a lot by having had the opportunity to work with him. His insights and guidance have been crucial to the success of this work.

This report is based on a Ph.D. thesis submitted to the University of Illinois at Urbana-Champaign in September of 1981 ([Baker 1981b]).

[1] increasingly so this past year

[2] and getting younger

[3] especially

## TABLE OF CONTENTS

# PREFACE

## *Abstract*

*The past few years have seen a growing interest in the application of three-dimensional image processing. With the increasing demand for 3-D spatial information for tasks of passive navigation, ~~([Gennery 1980], [Moravec 1980])~~, automatic surveillance, ~~([Henderson 1979])~~, aerial cartography, ~~([Kelly 1977], [Panton 1978])~~ and inspection in industrial automation, the importance of effective stereo analysis has been made quite clear. A particular challenge in this area is to provide reliable and accurate depth data for input to object or terrain modelling systems (such as ACRONYM [Brooks 1981a]). This report describes an algorithm for such stereo sensing. It is founded on an edge-based line-by-line stereo correspondence scheme — one which provides this three-dimensional analysis in a fast, robust, and parallel implementable way. Its processing consists of extracting edge descriptions of a pair of images, linking these edges to their nearest neighbors to obtain the connectivity structure of the images, matching the edge descriptions on the basis of local edge measures, and cooperatively removing those edge pairings formed by the correspondence process which violate the connectivity structure of the two images. A further matching process, using a technique similar to that used for the edges, is done on the image intensity values within intervals defined by the edge correspondence. The result of the processing is a full image array depth map of the scene viewed.*

## *Organization of this Report*

Chapter 1 discusses some of the psychological and neurophysiological aspects of the human vision system that have had an impact on this work, and within this context lays a basis for the direction of the research carried out. The fact that the research is being developed for implementation on a sequential machine, rather than a parallel mechanism as in the human system, imposes (or allows, depending upon the particular benefits/deficiencies perceived) certain constraints on the techniques used. Despite this distinction in the mechanisms available, the philosophy of the approach taken here has, at an informational level, strong parallels to the human system.

Chapter 2 outlines the main differences between the two principal techniques for binocular stereo analysis — those based on cross-correlation of image intensity values, and those working with image intensity contours, or *edges*. The functioning of the principal exemplar systems from each of these areas is described, and comments on these provide a background for specifying the goals of this research. Although providing a good summary of the state-of-the-art in stereo matching, Sections 2.1 and 2.2 rather go on, and a casual reader, looking for the meat, would be best advised to skip them. The chapter ends with a summary overview of the composite algorithm developed in this research. The algorithm, as the title of this report indicates, incorporates both edge-based and intensity-based analyses.

Chapter 3 introduces the principal unit of the analysis, directional zero-crossings in the second difference of image intensity, and identifies the particular geometric and photometric constraints that are integral to the analysis.

Chapter 4, dealing with the statistical measures used throughout the analysis, should be read in conjunction with Chapter 5, which discusses the Viterbi correspondence algorithm and the modification to it developed for the matching required here. The two chapters work together to define the matching process, with the first giving the hairy details of the decision metrics for the various correspondence processes, and the second showing the way these enter into the computation. Chapter 5 ends with a full exampl. of a single line-pair *edge* and *intensity* correlation.

Chapter 6 presents a cooperative algorithm which enforces global consistency on the tentative edge matches proposed by the preliminary analysis. Its presumption is that connectivity in the two-dimensional projection of a scene is indicative of continuity in the 3-space of the scene.

Chapter 7 provides two examples of the full processing of the algorithm — the first on a synthetic stereo pair of an urban scene, the second on real imagery of natural terrain. The success of the stereo matching algorithm on these images indicates that it is a powerful technique with general applicability; its failures suggest areas for future refinement.

Chapter 8 discusses the contributions of this research to the field of computer stereo vision, and highlights the areas where its application would bring immediate benefit.

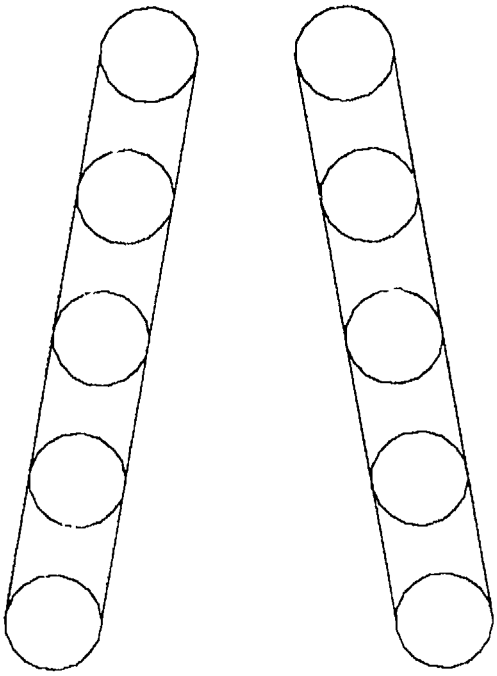
## *Viewing the Figures*

All of the paired figures in this report are drawn for *cross-eyed stereo fusion*. This means that to obtain the proper perception you must have the left eye see the right figure and the right eye see the left figure. The difficulty with so configuring your eyes is that you have probably never in your life before consciously decoupled your focus from vergence. Your eyes may be aimed in the proper directions, but since this attitude corresponds to the normal eye position for examining the tip of your nose, the focus is set at about that distance. For stereoscopic fusion in this situation one must consciously vary the focus while maintaining the fixation until the desired image is seen clearly. It will take a while if you haven't done it before, perhaps a half hour for a dedicated attempt, although once attained you will surely (it is my hope) find the effort worthwhile.

The choices possible in stereoscopic presentation are divided between those requiring special viewing aids, these demand little of the viewer, and free fusion techniques, which require no aiding devices but at the cost of considerable initial effort for the viewer. Anaglyphic depictions, where the two images are presented in complementary colours and must be viewed with suitably chosen complementary filters, are likely the most familiar to you. Another technique for slide/film presentation is to polarise oppositely the light passing through the two images and provide polarised filters to the viewers to ensure delivery of the proper image to the proper eye. Neither of these techniques is suitable for standard xerographic reproduction.

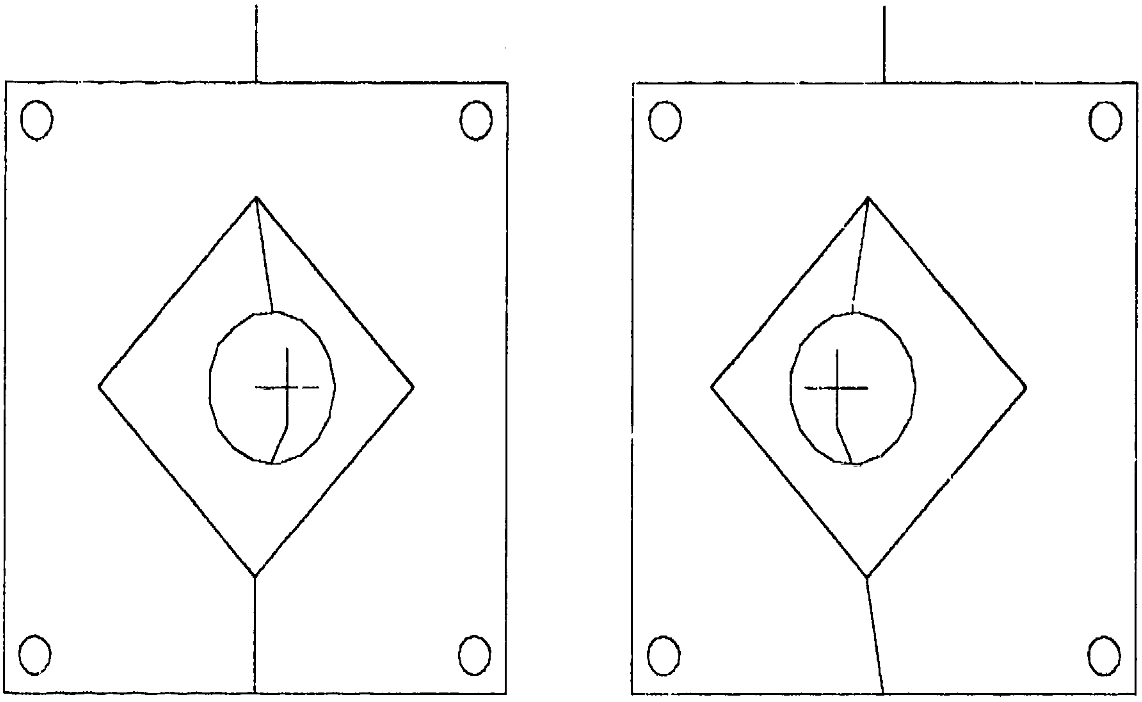
The technique chosen here, *cross-eyed stereo fusion* has one principal advantage over the other form of unassisted fusion, often called *wall-eyed fusion* or simply *free fusion*. In *wall-eyed* fusion the figures are presented so that to obtain the percept the left eye must be directed toward the left image, and the right eye be directed toward the right image. Since only in exceptional (and then likely damaged) systems can the eyes actually diverge, the separation between the centers of the two images cannot, in general, exceed the interpupilary distance of roughly 7 centimeters. This means that the figures can be no wider than about 7 centimeters — a clear handicap when using limited resolution graphic devices. There is no such limitation for *cross-eyed* fusion. Its advantage is then clear — resolution of depictions can be much greater and figures may be projected onto distant screens for audience viewing. (Oddly enough, the stereoscopic perceptions do differ in these two cases, with wall-eyed fusion giving a greater sensation of relative depth for the equivalent monocular percepts. Vergence, not just observed disparity, affects the judgements of stereopsis.)

Figures 1 and 2 contain exercises which may be of some help in enabling you to develop the skill (skill?) of cross-eyed stereoscopic fusion. When working with them, remember to have the left eye fixating on the right figure and the right eye fixating on the left figure. The captions to figures give suggestions for their viewing. If you can find a pair of 2- or 3-power telescopes and aim them cross-eyed at the appropriate images, this will greatly ease the task of obtaining stereo fusion. By reducing the effective focal distance, the lenses will compensate for the vergence/focus coupling conflict, and should allow the images to be more readily fused. But it is better to avoid reliance on such an aid — you may find yourself somewhere, sometime, wishing to fuse two disparate images and have your aid nowhere at hand. Unequipped, you will have missed your chance! (This generalizes.) Figure 2-5 seems to be a good example for practise in cross-eyed fusion.



First, fixate on the two circles at the top (if necessary, keeping the lower circles from distracting you by covering them with a piece of paper). When you have superpositioned these so that the left eye and right eye together see a single (probably blurred) blob, try to focus the eyes to make the images clear. If you can do this (this is the single biggest problem in fusion — controlling the focus), then slowly move your gase down toward the next pair of circles (following the connecting lines). The percept should be of continually progressing circles, forming a tapered cylinder (an interesting visual illusion in itself).

Figure 1



This figure may be more difficult to fuse. Begin by bringing the outer squares into alignment – the top line flash and lower left and upper right circles may be helpful in controlling the vergence movements to bring the images together. Once superpositioned, work on obtaining clear focus, perhaps by concentrating on one of the corner circles or the upper flash. Once you have this, follow the box perimeter around to the lower line joining the box to the diamond, follow the diamond perimeter to the circle, and then on to the cross. You should see the cross lying farthest away, and the framing box nearest.

Figure 2

Here is yet another trick to try if neither of these figures seems fusable even after hours of hopeless staring. Cut a window about 1.75 inches square from a piece of cardboard. Hold the images about 15 inches away from your nose, the card about 5, and line up the two images through the hole so that the left eye sees all of the right image and the right eye sees all of the left image. If you concentrate on looking at the frame of the window (not at the scene beyond quite yet!) you will get a vague impression of the intended depth. Work at keeping your regard on the frame while gradually letting your focus slip through the two windows to the images below. It shouldn't be too long before you are able to separate the focus from the vergence and see the three-dimensional scene below with whatever your customary clarity.

# *Chapter 1*

# CONTEXT OF THE RESEARCH

Let me draw your attention, at the beginning of this discussion, to the phenomenon of man's stereoscopic binocular vision — the fusion of the left and right eyes' images into a coherent perception of three-dimensional space. This single and immediate perception of the dimensionality of our world is a striking achievement. To that minority among us lacking binocular stereoscopic perception (at least 3%, while as many as 15% may have stereopsis deficiencies ([Bishop 1975], [Richards 1970])), it is an experience impossible to describe by analogy. It is unique in character, likened in its vividness to the perception of colour. This visual system, called *"the most intricate structure in the known universe"* ([Julesz 1976]), has been one of the principal contributors to our species' intellectual and sociological development. An increasing dominance over time of the visual sense has, through its interaction with manipulative skills, enabled us to become the best living tool users, constructors who have the ability to mold the world around us to our needs.

An important consideration in the implementation of a machine vision system is the impact knowledge of this marvellous human system should have on the machine system's design. To know something of its development, functioning, and mechanisms would seem to be a prerequisite for a proper attempt at developing something similar for a machine (be the similarity in mechanism or effect).

## *1.1 The Stereopsis Process in Man*

In the course of primate mammalian evolution there has been a gradual movement of the eyes from a lateral-looking attitude to a frontal binocular position. This transformation enabled a considerable overlap in the visual fields of the two eyes — a necessity for stereo vision — and facilitated a precisely registered coordination of binocular eye movement. The development of a centrally located high precision fovea greatly aided this evolution of coordinated eye movement. These changes, and the neurophysiological developments in the cortex accompanying them[4] were also correlated with the development of hand-eye coordination. The adoption of the upright attitude freed the hands from their previous role in postural support, and enabled the development of manipulative skills under visual (especially foveal) guidance.

Precision in depth determination is one of the principal advantages of stereopsis — it allows accurate hand-eye functioning and visual tracking. Our stereo acuity has been estimated as being between 14 and 40 seconds of arc for normal bifoveolar binocular vision ([Bishop 1975]); the binocular luminance threshold has been found to be as much as a factor of $\sqrt{2}$ lower than the monocular threshold, and visual acuity increases accordingly for binocular over monocular vision. These observations bear out the statistical improvement expected from using two independent measuring processes, the two eyes, rather than one. Beyond the statistical advantages are those of more practical importance — stereopsis enabled us predators to *"see through"* the camouflage by which hunted animals sought to blend in with their surroundings[5](monocular camouflage fails to a stereo perceiver in the range over

[4] development of a partial decussation at the optic chiasma and the organisation in the visual cortex, having corresponding fibers from the two retinas synapsing on the same cells in the striate cortex

[5] Our prey had the complementary advantage of nearly complete peripheral (360°) vision to see us coming.

which stereo is effective — theoretically, up to about 500 meters[6]([Bishop 1975])), and it enables us a perception of solidness in our visual worlds isomorphic to the solidness of our physical world.

The eyes are positioned about 7 centimeters apart in the head, observing the external world through central projections from two slightly differing viewpoints. This difference in position causes varying relative lateral displacements, or disparities, of the detail in the two images projecting onto the retinas. It is these disparities, the differences in horizontal position of corresponding points on the two retinas (varying directly with the distance from the point of fixation), which provide the essential data for binocular depth perception. The ***fusion*** of the two retinal images into a single perception of solid 3-space, the process of binocular stereoscopic vision, is termed ***stereopsis***. The subject of this report is an automating of this process of ***stereopsis***.

## 1.2 Computer Vision

The involvement of computer scientists with visual processing is in the use of computers as sensory data processors for observing and manipulating the environment. Generally, the interest is in bringing the control advantages of visual sensing to the tasks of robotic manipulation and autonomous navigation. It might be thought that mechanisms chosen for this would be selected more on their algorithmic tractibility than on their relevance to neurophysiological or psychological theories of human perception. This is true to an extent. We have access to neither the parallel processing biological mechanism that resides in man nor an adequate definition of its functioning, and are forced largely to rely on sequential machines for the implementation, and introspective insights for guidance in our algorithms. Still, human visual functioning is our principal source of observations on the process of three-dimensional vision, and it supplies the best paradigm we have for a seeing system. Clearly, where insights from human visual processing would add to the robustness or flexibility of the system, our machine should have them. On a more pragmatic note, it makes excellent sense to pay attention to human visual functioning, for it provides the best of insurance; the problem is solvable, and the human system does it.

In the present work we do not aim explicitly for our algorithmic system to have biological feasibility, but we do wish to have it parallel the highly effective functioning of the human system — a functioning where the input is passively sensed (although perhaps actively pursued) visible light, and the result is an ***understanding*** of the physical environment.

There will be obvious hindrances to our work — the computers available are only serial devices (at the moment), and the mechanisms of sight are little enough agreed upon by neurophysiologists and perceptual psychologists; neither our devices nor our algorithms can yet approach the power and flexibility of the human system. But the goal is there, to develop mechanized vision. Although our computing devices are not ideal for the job, they are adequate for the research; although our understanding of the process is partial, we have sufficient empirical observation to allow us reasonable insight into the operation of our sight.

On this last point, it is interesting to note that computer vision is, increasingly, developing a symbiotic relationship with studies in human perception:

- to implement a theory requires complete and detailed specification of the process — this invokes a rigour at the level of the definition,
- as an experimental tool, the computer stimulates insights which improve the theory.

[6] the greatest distance at which an object can just be detected as nearer than an object at infinity

So, while machine vision researchers look to the perceptual psychology and neurophysiology literature for insight into mechanisms and measures of adequacy for their algorithms, perceptual theorists turn to computers as tools for their studies and means for instantiating and testing their theories ([Marr 1976], [Marr 1977], [Mayhew 1981]).

## *1.3 Considerations for a Stereopsis Process*

### *1.3.1 – Possible mechanisms of human stereopsis*

What is involved in human vision in going from the sensory stimulation of the two separate retinas to achievement of the depth understanding of stereopsis? Perhaps one or some combination of the following:

1) the independent monocular recognition of each eye's contents, and a subsequent *matching* of recognized items across eyes for distance determination.
2) a less *knowledge-intensive* matching process, whereby *features* (perhaps 'blobs') characterized by uniformity of some property are extracted from each eye's image and compared across eyes (without any familiarity with the particular *features*).
3) extraction of some information-specific abstraction from the images of the two eyes (for example, zero-crossings in the second difference of a laterally inhibited signal), and the matching of these sampled items across eyes.
4) matching of individual brightness levels over the entire images of both eyes.

The distinctions between these lie in the level of *abstraction* attained. The actual monocular recognition of scene content is a great abstraction — image brightness values are clustered to define shapes which are representable as symbolic descriptors — whereas the matching of brightness levels, being little more than photon counting, can hardly be considered as abstraction at all.

The first suggestion above requires the process to have a monocular familiarity with everything in the scene, implies that whatever it is, it can be recognized when viewed from any perspective, and grants of the monocular processing a quite remarkable capacity at separating objects from each other and from their background. With this scenario, the eyes work independently up to the point of placing the depth component on the object's position. Alone, this is not a very satisfactory explanation of stereo perception. It presupposes an unsubstantiated *inner eye* projection system for mapping 3-D *known* forms to percepts and, most significantly, provides no mechanism for the *learning* of new objects. It is contradicted by known characteristics of the human visual system's processing in its *sufficiency*, by our ability to fuse random-dot stereograms ([Julesz 1971]), and in its *precision*, by the accuracy and continuity of our depth perception. It is presented here as a straw man, merely to focus on this extreme of stereopsis possibilities.

The second suggestion is an improvement over the first, in that it demands no monocular object recognition, yet it still hinges on the ability to extract information that is *meaningful* across images. Considered as the **sole** process for stereopsis, it has inadequacies similar to those of the first suggestion.

Implicit in the discussion of suggestion 2 was that it dealt with a *uniformity* measure on the imagery. Consider suggestion 3 as involving the processing of a *discontinuity* measure. The sort of information-specific abstraction suggested can vary from individual *edge* elements to extended contours, perhaps delineating the outline of some shape (see [Wilson 1978a, Wilson 1978b] for a discussion of spatial frequency filtering in human vision).

Suggestions 3 and 4 are similar to each other in that neither presupposes a semantic processing of the retinal images and both involve extensive cross-correlation on a great quantity cf data. Human stereopsis supports approach 3 more strongly ([Wilson 1978a], [Marr 1979], [Schumer 1979]), while not excluding the possibility of interaction with an intensity matching process[7] as suggested in 4. For a machine implementation, questions of computational cost, viewing constraints, reliability, and desired accuracy will affect the utility of one over the other, and this will be discussed further later.

Observations of visual processing make it evident that, when impoverished, human perception can rely upon most any of the above techniques for depth determination.[8] None of them is sufficient for visual understanding; beyond each must lie a process bringing a unity of interpretation to the measures. Monocular processing can aid stereopsis by establishing a context or vergence setting ([Saye 1975]), and can enable fusion despite conflicting evidence at a local level (as demonstrated by Helmholtz (1906) with positive image/negative image fusion (see page 157 of [Julesz 1971])). Binocular stereoscopic processing can reveal depth when no cues are available to the eyes in isolation. Psychophysical evidence ([Gregory 1977]) suggests that the monocular and "*cyclopean*" processes ([Julesz 1971]) may be highly independent functions.

### 1.3.2 – Primary versus secondary cueing for stereopsis

Suggestions 3 and 4 use what are termed *primary* cues for stereopsis — information that relies on analysis from both eyes working in unison. This *primary* stereopsis is immediate in the sense that it provides local depth information everywhere obtainable in the visual field, information that is unavailable from the eyes individually. It might be said that the percept occurs before cognitive influences can play a role. Complementary to this is the use of *secondary* cues for visual depth perception. Our species has learned much about the environment we have lived in over the past few millenia that greatly facilitates the making of subjective visual judgements -- judgements that can be made on the basis of information presented to either eye, independently. These are monocular depth cues. Fallible as they are, such cues (see [Gibson 1950]) as:

- object overlay or partial occlusion,
- perspective deformation,
- brightness and shading,
- texture density gradient,
- motion parallax,
- hue variation, and
- object relative size,

provide for remarkable judgements of relative depth from a single monocular view. There is no doubt that these cues, irrespective of their classification as *secondary*, are principal contributors to modern man's perception of his world.

An important point to note here is that *secondary* cues to stereopsis contribute an explicit *globality* — they have a spatial component that relates them to parts of the visual field in their locale. A similar provision is implicit in *primary* stereopsis in that local correspondences (depth judgements) interact to produce the optimal percept for a stereo pair of views — a more global analysis is at work

[7] although issues of optic nerve bandwidth will impact upon this possibility — it seems unlikely that all photometric information is transmitted along the optic nerve to the lateral geniculate body. There are roughly $10^8$ rods, $10^7$ cones, and, with a mere $10^6$ optic nerve fibres, substantial coding would be necessary.

[8] There is no psychophysical evidence that I have seen supporting the first suggestion.

to ensure some form of continuity or consistency in the three-dimensional interpretation. Automated stereopsis must also be globally consistent.

It is hazardous to argue about the evolutionary *development* of man's depth perception — as to whether *primary* analysis preceded *secondary* analysis, or which has been the dominant factor in our visual development. Clearly if our visual perception progressed from the lateral-looking attitudes of our presumed genetic ancestors, then we may have had some form of strictly monocular processing (perhaps with temporal stereo) before the occasion arose to try any fusion, so the monocular analysis would have had a head-start on affecting our development; yet certain *secondary* cues have been determined to be consequences of experience (for example perspective deformation, as demonstrated by the Ames room phenomenon and the experiments of [Yonas 1978], and texture variation in the visual cliff experiments of Walk and Gibson [Walk 1961]), so are conceivably learned — it would appear that they are mediated by higher-level functioning.

From an implementation standpoint, the choice of approach is one based on sufficiency: it seems obvious, at least to the author, that the *primary* sensing mode can provide information to allow the development of a *secondary* cueing mechanism, while the reverse does not seem to be true. From this perspective, *primary* stereopsis is the most interesting. In truth, the two mechanisms probably developed separately, and exist independently — although cooperatively — in constituting our vision system.

### 1.3.3 – The necessity of a primary cueing mechanism

Useful as they are, depth estimates based on secondary cues do not have the same perceptual quality and accuracy as do those due to stereopsis. Secondary cues provide a cursory form of processing. They may be seen as arising from abstraction over time of the information presented by the primary stereopsis process. Our capability at attaining a perception of depth relying solely on the primary stereopsis process is well documented. The easily-learned fusion of Julesz' random dot stereograms — which have neither monocular depth cues nor monocular structure — is a convincing demonstration that stereopsis is at work in our visual system. Under circumstances of contextual deprivation, stereopsis enables the perception of depth.

This argues that for a machine approach to vision a dominant consideration should be in specifying a stereopsis process — one which autonomously, and without the aid of domain-specific or environment-induced knowledge, constructs a depth map of the field of view. The contribution of this report is in a definition and demonstration of a domain-independent stereo correspondence algorithm, one which can use certain monocular cues where available for ambiguity resolution, but functions in the *primary* binocular mode in attaining the depth determinations of stereopsis. The *philosophy* underlying the design to be presented here will hopefully be seen as having some relevance to human visual processing, although the *mechanisms* developed for the computation will be chosen strictly for their effectiveness and efficiency as implemented in a serial machine. (Although, as will be seen later, the *structure* of the computation has been chosen so as to facilitate a parallel implementation.)

# *Chapter 2*

# BACKGROUND TO MACHINE STEREO VISION

## *2.1 Area-Based Versus Edge-Based Processing*

Much of early machine vision work avoided the aspect of three dimensionality inherent in man's perception of his environment, and relied upon projective monocular measures for its analysis of visual domains. In the last eight to ten years, though, there has been a growing strong interest in three-dimensional sensing and analysis, and this has brought with it several differing approaches to the problem of matching the content of a stereo pair of images. The primary division among these research efforts is one of *area-based* versus *feature-based* analysis.

The distinction between 'feature' and 'area' correspondence here can be more a matter of degree than type. *Feature-based* analysis has involved the transformation of the sensed data from a discrete two-dimensional intensity array to a more symbolic form as significant intensity contours, or 'edges' - features. It is the properties of these features which then provide the metric for the correspondence. 'Feature' is a fairly general term, but its use here may be equated with 'edge'. There are many fewer 'edges' than image elements in a view of a scene, so this transformation, generally, reduces the computational cost of determining correspondences. A corollary, and noticeable drawback of this, is that not every point in an image is a 'feature', so the result of a solely feature-oriented correlation will not be the dense depth map one may want.

### *2.1.1 – Area-based analysis*

In *area-based* analysis two-dimensional windowing operators measure the similarity in intensity pattern between local areas, or windows, in the two images. Cross-correlation is used to determine matches between windows in one image with windows in the other. *Normalized* cross-correlation has the ability to compensate for contrast and brightness differences across images. If the lighting and sensor/processing conditions are known, this flexibility in the algorithm may not be required. In this case other correlation forms such as *Normalized RMS* or *Absolute Difference* may be used (see [Hannah 1974] for a summary of these differing techniques).

Area cross-correlation is often not applied to every pixel in the image arrays, but selectively for those whose local variance is high. With this approach, the variance measure is used as a filter to limit possible correspondences; correlation is then used to select the best from among the candidates. These variations may qualify such approaches as 'feature-based', although they will not be considered so here. Perhaps a better way of categorizing these systems is as *feature-driven area-based.* [Levine 1973] limits initial correlation to areas having local maximal variance, [Henderson 1979] preprocesses the data to find edges which are then used to bound an area-based search, [Moravec 1980] uses an 'interest operator' to select significant points in a reference image, and [Gennery 1980] uses a variance based $F$ test to filter out areas of minimal information, and therefore minimal interest.

Area-based correspondence has been applied quite successfully to the stereo analysis of rolling terrain, but it degrades when the scene is not smoothly varying and continuous. In images of such domains many windows to be matched will have no correspondences in the other image (for example, those windows lying on surfaces which are occluded from the other imaging position). The chief difficulty with the area-based approach is in properly matching window shapes and sizes for conjugate image areas, taking into account both variation in terrain slope and discontinuities at surface boundaries (see [Ryan 1979] and [Ryan 1980]).

Large correlation window sizes are required in attaining statistical significance in the sampling, yet the characteristics measured over the windows become less and less representative of the observed local surface as this window size increases. Discontinuities in the surface can cause a positioned window in one image to be sampling local intensity values from more than one intensity surface in the other image, and a correct cross-correlation would only be possible if the window could be partitioned and matched with (possibly several) windows of various size and shape in the other image. Such adaptation requires more flexibility than area-based correspondence has thus far been shown to provide. Abrupt discontinuities in topographic structure and an abundance of occlusions characterize urban or cultural areas. It is at precisely these points of depth discontinuity that we want to obtain accurate surface position measurements. This would suggest that current area-based processing is inappropriate for domains with occlusions and abrupt depth discontinities.

Some consideration of this window shaping problem has been attempted in area-based work. [Levine 1973] and [Mori 1973] vary their correlation window sizes with the local intensity variance. They presume that high variance implies high local texture and thus suggests the need for smaller correlation windows, while low variance suggests surface uniformity and the need for larger sample sizes and larger correlation windows. [Panton 1978] uses trapezoidal window shapes in the search image, as determined by previous and predicted correspondence results, to match the rectangular windows of the reference image. [Gennery 1980] included a partial solution to this problem for a specific camera geometry when looking at windows presumed to lie in the ground plane. [Mori 1973] implemented an iterative technique that would compensate for terrain variations by successive refinements to image registration estimates. Both [Levine 1973] and [Hannah 1974] included in their algorithms techniques for identifying certain scene occlusions and areas of image non-overlap, but these were entered more as cases of exception handling, and it is doubtful that they were adequate as models of occlusion.

A related problem with area-based correspondence is that increasing window size improves statistical significance but generally results in poorer 3-space positioning accuracy for the correspondence. Feature-based analysis obtains more precise positioning (for its edges) in the individual images, and it can attain correspondingly higher accuracy for its correspondences in 3-space ([Arnold 1978] indicates that edge-based techniques offer an order of magnitude improvement in accuracy over area-based correlation methods).

Area-based correspondence systems also tend to be prediction driven, in that they process an image serially and at each step use the context of previously matched neighbouring points to limit the search for the present correspondence. None provides a backtracking facility with this technique, and only [Gennery 1980] includes a scheme for adjusting locally determined miscorrespondences. With little ability to either correct or detect errors, such prediction-guided approaches can lead to rapid degeneration once errors begin to occur.

A final and important anomaly to note of area-based processing lies in the basic philosophy of its analysis. The underlying assumption of area-based correspondence is that it is the *photometric* properties of a scene that are invariant to imaging position, and the correlating of these properties will be sufficient to allow the proper correspondences to be determined. But it is not the measurable photometric properties that are invariant to viewpoint positioning. In the degenerate, although common enough case, a surface of a certain intensity seen unobscured from one viewpoint will not even be visible from another slightly different viewpoint. All that can be said to be *truly* invariant to viewpoint positioning is the *three-dimensional structure of the scene itself.* A better metric for the correlation would be one which deals in some way with that scene three-dimensional structure. I will return to this point in the discussion of feature-based correspondence methods.

### *2.1.2 – Area-based correspondence methods*

Mapping systems available commercially, and used in the photogrammetry community, are exclusively *area-based* in their analyses. State-of-the-art photomapping devices employing automated correlation include the Bendix AS-11B-X ([Scarano 1976]) and the Gestalt GPM-II ([Kelly 1977]). These systems are not, in general, of much interest algorithmically; they have inadequate success rates for the correspondences they produce (failing to determine scene depth at between 40% and 70% of image positions, according to recent studies, see [Friedman 1980]), and require extensive manual intervention for their operation. More fruitful insight to the potential of cross-correlation techniques can be obtained by looking at systems produced in research, rather than development, environments.

The following summaries describe the more important area-based stereo correlation research systems of the past decade. The last three are the most recent and most accomplished of these systems.

#### *Gimel'farb, Marchenko and Rybak System 1972*

[Gimel'farb 1972] was the first report to document the use of dynamic programming[9] in a stereo correspondence process. The algorithm described processes image pairs on a line-by-line basis, exploiting epipolar geometry constraints and using known (*a priori*) disparity and surface slope limits to constrain the correspondence search. It optimizes a cost function of normalized cross-correlation. The convolution incorporates a lateral inhibition computation. The correspondence algorithm is described analytically as finding the function mapping intensities from one image line to the other. Testing was done on short wide images (*i.e.* 5x500). The authors suggest that one could improve the speed of such stereo processing in two ways. First, in using the results of prior line analyses to guide the matching and bound the search on subsequent lines, and second, in partitioning lines into smaller stretches, reducing the combinatorics of the correspondence matching. The first is a technique that CDC used in their stereo work (as will be discussed). The second can be seen as a preview of the multiple resolution correspondence processes of Baker, when it is seen that rough alignment of corresponding parts of the two lines must be made before breaking them into smaller stretches. Depiction of the results obtained with the algorithm are a bit sketchy, as the plots shown are of single line analysis only. The report comments that results from this totally automated process are comparable to those of human operators using automated photomapping devices, although nothing quantitative is presented.

#### *Levine and O'Handley System 1973*

[Levine 1973] describes a system designed to provide depth information for the Mars rover vehicle's autonomous navigation. Tests of its performance were carried out on stereo imagery collected in the vicinity of the Jet Propulsion Laboratory. Because of the system's intended use, it was possible to work with the basic premise that the scene viewed was approximately planar, running off to a horizon somewhere in the distance (not necessarily in the images). It used collinear epipolar imaging[10]for its two cameras to limit correspondence search. Matching was by intensity cross-correlation, with an adaptive window size set by the variance at pixel $(i,j)$ in the image – a large variance sets a small window size, and vice versa. Processing was organized to run in lines from bottom to top. Search constraints on possible disparities were exploited throughout the analysis. First the top and bottom lines were correlated to estimate the overall disparity ranges (notice that this presupposes that scene depth varies regularly from top to bottom, as in a view toward the horizon). Then a

[9] see chapter 5 for a discussion of this

[10] see section 3.2.1 for a description of collinear epipolar geometry

prepass analysis was applied to a sampling of $n$ lines ($n = 5$) to set local maximum and minimum disparity ranges. Correlation along a line pair was over windows with locally maximal variance, called 'tie-points'. The local maxima were used to iteratively segment the reference line. A coarse search using statistical parameters (variance) of image windows was used to find good candidates for the more expensive computation of the correlation coefficients. The candidate pairings chosen through this process were then evaluated to select the optimal matches and to refine their positions in three space. The coarse search was done with every other pixel along a line. Cross-correlation was only done with windows of similar variance. The system uses the epipolar geometry constraint in a way that prohibits positional reversals along a line. The authors indicate in the paper that they are aware of the difficulties introduced by occlusions, and mention an ad hoc scheme for preventing parts of the images felt to be occluded from being matched, but the technique is not further described. Two-dimensional proximity was also used to limit disparity possibilities; an allowable range was set at each tie-point by examining neighbouring disparity values on the preceding line (actually the current line minus 4 — *i.e.* they process every fourth row and every second element). Final disparity values were smoothed, and deviants removed.

### *Mori, Kidode and Asada System 1973*

Mori, Kidode and Asada, in a short paper [Mori 1973], describe an interesting stereo mapping system. In it, epipolar geometry is used to constrain the search for correspondences in the area-based correlation they use. The system is demonstrated on a model pattern and a pair of aerial photographs, although only a single line of results is presented. A gaussian weighted correlation function is used to diminish the effect of peripheral intensity variations. Window size is modified by the range of disparity expected for the point, and they suggest that this should be set by first correlating over a large window, then narrowing to a smaller window when the gross disparities are known (the paper doesn't explain this resolution reduction process any further). An assumption of scene continuity is also used in limiting correspondence search. The technique is iterative: the right image is repeatedly distorted and compared with the left image until no substantial intensity differences are found. The abstract says that the first matching is done on highly contrasting parts of the images ('roads, coast, forest edges'), and the context of this is used, with the smoothness assumption, to expand the correspondences into neighboring parts of the scene; but the body of the paper does not elaborate on this. The paper is very brief and cursory, suggesting much more than it reveals. It would be very interesting to see whatever further documentation they have on this system. Examples are incomplete and inconclusive. No follow-up has occurred to this work.

### *Hannah System 1974*

[Hannah 1974] describes a series of techniques developed for increasing the efficiency of area-based correlators. Her thesis contains a discussion of the differences between Discrete Correlation, Normalized Cross-Correlation, Normalized RMS Correlation, and Normalized Absolute Difference. The work takes an experimental approach, and documents the improvements arising from:

- correlating over a sampling of the image arrays, then refining the match estimate using the full arrays at the point having maximum correlation coefficient (this is referred to as 'gridding'),
- correlating over reduced resolution depictions of the images, and then refining match estimates with the higher resolution depictions,
- abstracting area characteristics (mean/variance), and using these more symbolic descriptions for limiting windows to be cross-correlated,
- using known camera geometry constraints to limit search.

A region growing approach is taken in expanding correspondences outward from matched pairs (using an assumption of surface continuity). Various heuristics are introduced for inferring the distinctions between occlusions, corrrespondence errors, and out-of-scene overlaps. Hannah introduced here, through the autocorrelation function, a means of assessing the quality of area-based matches.

*Panton System 1978*

Panton's paper [Panton 1978] describes a system for obtaining a dense digital depth map of smoothly rolling terrain. The algorithm, using intensity cross-correlation, processes from left to right in the images, and so, once initialized, can use local context of previous matches and estimates based on the epipolar geometry to provide tight constraints on possible correspondences. Maximization of a correlation coefficient in the chosen area selects the appropriate match. About 1% of the pixels in an image are matched in this manner, although the entire image is used in determining the match correlation coefficients. Positioning accuracy of somewhat better than one pixel is obtained. The system is able to tailor sampling window shape in one image to follow roughly the deformation of the rectangular window it matches in the other image. This window-shaping issue is one of the principal difficulties of cross-correlation analysis — only in the case of flat terrain normal to the line of sight are corresponding windows in the two images of the same shape. Panton's solution to this is to approximate the rectangular source window by a trapezoidal window in the other image. The technique is based on a large sampling of the surrounding neighbourhood, and uses the terrain relief predicted by previous neighbouring correspondences to estimate the shape of the trapezoid about a candidate surface point. Trapezoidal shaping is quite an improvement over matched windows, but is still just an approximation to the actual projective situation. This algorithm has been implemented in an experimental parallel processing machine which seems to achieve quite impressive performance in processing on relatively smooth natural terrain. It is not clear whether or how much operator intervention is required.

*Moravec System 1980*

Moravec's research (see [Moravec 1980]) was aimed at providing vehicle control information from visual sensing. His aim was not to construct a depth map, but rather to sample interesting points in a scene and use these to provide motion calibration information and obstacle cues. There are three main vision contributions in his research: the ***interest operator***, the ***binary correlator***, and ***slider stereo***, the first two of which have been widely adopted by researchers in the field. The ***interest operator*** and ***binary correlator*** date to 1974. The ***interest operator*** is a filtering technique for selecting points at the center of locally maximal directional variance — these are typically corners. The ***binary correlator*** finds the best match of a feature in one image with the intensities in the other image using a resolution varying technique. Each feature (as found by the ***interest operator***) is represented as a series of (5) 6 × 6 windows, in increasing resolution (*i.e.* 6 × 6, 12 × 12, 24 × 24,... in the original image). The lowest resolution description of the feature from the reference image is moved a pixel at a time over the other reduced image, calculating correlation coefficients at each location. The largest correlation coefficient is taken as indicating the best match. The next higher resolution window (*i.e.* next smaller window) centered on this is then searched (with the next higher resolution of the feature). This correlation process continues until a 6 × 6 patch is matched in the unreduced image. The correlation has about a 10% error rate. In ***slider stereo***, lateral movement of a camera along a track provides 9 equally spaced camera stations. Correlation of the resulting 36 (9 choose 2) possible image pairings provide a series of estimates of distances to scene points. These estimates are represented as gaussian distributions (mean equal to the distance estimate, and the standard deviation inversely proportional to the baseline) weighted by the correlation coefficient of the feature matches (from the binary correlator). The 36 histograms (distributions) are then summed, and the peak taken to indicate the correct match. Stereo tracking between vehicle motions is also performed with the ***interest operator/binary correlator*** techniques. Here, features from the central image at the previous position are searched for in the central image of the current position, and the results of this correlation inform the system of the vehicle's actual movement. The positional and depth information obtained from these correlations provide data for the navigational control of the vehicle. It knows roughly how far it has moved through the scene, and where its obstacles lie. Feature sampling is chosen so as to cover most of the scene, uniformly.

### *Gennery System 1980*

Gennery's system [Gennery 1980] was designed to provide depth data for vehicular autonomous navigation. It uses cross-correlation to position points in space. The system incorporates a ground plane finder (utilizing Moravec's ***interest operator*** and ***binary correlator*** [Moravec 1980]) that estimates a plane in the scene above which most points lie, and uses this to estimate the camera relative orientations. This derived camera relative orientation information enables the matching of corresponding windows to be constrained to a one-dimensional search. Having estimates of scene noise characteristics (variance, and gain and bias between the two images), he defines a correlation measure that provides sub-pixel positioning of corresponding windows. Accompanying these are estimates of the confidence and accuracy of the correspondences. Since it progresses across an image from left to right, his algorithm can use local context of previous matches to suggest tentative match sites. If these are inadequate for unambiguous matching of the particular window, search constraints based on the epipolar geometry can be used to provide further suggestions for the correspondence. These begin at the *infinity* point of the corresponding epipolar line (disparity equals zero), and come forward (to the left, with increasing disparity) until either a suitable correspondence is found or some already matched windows are encountered. When the correct locale has been chosen, maximization of a correlation coefficient in a vicinity of the selected area determines the local best match. This analysis is followed by a process of fitting ellipsoids to the determined elevation data. These, he contends, are an appropriate shape representation for use in obstacle avoidance calculations and scene matching.

## *2.1.3 – Feature-based analysis*

Recall that area-based analysis was criticized as being based on a metric sensitive to imaging position. Feature-based analysis avoids much of this problem, and comes closer to dealing with the true invariant of the projection process: scene structure. It works generally with the premise that a local measure on the intensity function is representative of physical change in the underlying scene. The local measure on the intensity function could be, for example, a maximum in intensity gradient — peak in the first difference of intensity, zero-crossing in the second difference. Physical change in the scene could be a break in depth continuity and accompanying projected surface reflectance or luminance change, or a change in surface intensity from a surface detail without topographic break. The point to notice is that feature-based analysis uses the semantics of intensity variation in its attempt to extract measures of the physical change in the underlying structure of the projected views, and uses these two-dimensional observations to infer the three-dimensionality of the scene. The validity of this intensity edge tracking in a stereopsis system is apparent:

- a discontinuity in surface orientation will, in general, give rise to a variation in incident reflection which will appear to an imaging source as a change in brightness – tracking the intensity edge across the two views will track the surface discontinuity;

- a discontinuity in surface reflectance (a surface marking or pigment change) can be tracked to reveal the three-dimensional position of the variation on the bearing surface;

- an illumination discontinuity (shadow edge), although not likely corresponding to a surface discontinuity (the shadow will lie *on* the surface), will be visible as a brightness discontinuity – tracking the shadow edge across the two views will provide depth information about the shadow-bearing surface;

### 2.1.4 – Feature-based correspondence methods

Probably the most widely known edge-based stereo scheme to date is tha of Marr and Poggio ([Marr 1977]), as implemented in a computer program by Grimson (see the following summary [Grimson 1980]). The algorithm has been fairly well tested on a reasonably wide variety of images (random dot stereograms, natural terrain, urban scenes), and is at present being implemented in hardware [Nishihara 1981]. [Arnold 1978] developed an edge-based stereo correspondence system that used local edge properties to select edge match possibilities, and a weighted iteration process to resolve match conflicts. The stereo processing system of Henderson, Miller and Grosch of the Control Data Corporation research group (as summarized in [Henderson 1979]), called the Automatic Planar Surface System, uses edges to guide it's area-based matching. They address their work specifically toward the problem of constructing planar models of rectilinear cultural structures from stereo pairs of aerial imagery. An extension of this CDC work ([Degryse 1980, Panton 1981]) has lead to a stereo matching system that uses both *local* edge information and *extended* edge information in its stereo matching. Some earlier systems whose simpler stereo processing was coupled with object modelling and recognition work will not be discussed here (for example, [Baumgart 1974], [Baker 1976], and [Burr 1977]).

#### *Arnold System 1978*

[Arnold 1978] describes an edge-based stereo correspondence system which uses edge orientation and side intensity, and edge adjacencies in determining the set of globally optimal edge matches. Examples are shown of the processing of aerial views of an aircraft, cars in a parking lot and an apartment complex. The Moravec interest operator and binary correlator [Moravec 1980] and a high resolution correlator and camera solver [Gennery 1980] are used in determining the relative orientations of the two imaging stations. The Hueckel operator [Hueckel 1971] is applied to the images, producing a set of edge elements for the correspondence. The derived camera attitude information is then used to reorient the edges to a canonic frame — one where the stereo baseline is along the $x$-axis and disparity shifts due to the tilt of the ground plane are cancelled. Disparities are restricted to those lying between zero (the ground) and some *a priori* limit in the $x$ direction. A list of possible matches in the right image is obtained for each edge in the left image. Loose thresholds are used to specify the adjacency structure of the edges. A reinforcement/inhibition voting scheme is applied to the adjacency structure and match list, and the resulting maxima are chosen as the correct matches. The technique uses many heuristics and thresholds, and is said to be quite sensitive to the output of the Hueckel operator.

#### *Control Data Corporation's Automatic Planar Surface System 1979*

The aim of this CDC work [Henderson 1979] was to provide automatic reference preparation capabilities; the references being structural models of buildings which may then, at a later point, be used in scene recognition for autonomous guidance. Because of this aim, they addressed their work specifically toward the problem of constructing planar models of rectilinear cultural scenes from aerial imagery. They took an interesting edge and area-based approach to their solution, using edge information to guide the application of a dynamic programming intensity correlation for line-by-line pixel matching. The principal contribution of their research is in this 'Broken Segment Matcher'. Roughly, their algorithm functions as follows:

- Geometrically transform a pair of images, bringing them into a collinear epipolar frame.
- Locate (via a Sobel operator) and 'thin' edges in the two images.
- Establish edge correspondences in the first pair of epipolar lines by hand.

- Maintain two cooperating correspondence processes to minimize the effects of image noise and extraneous detail. The first process matches intensities using only edges deemed to be 'reliable', such as those seeded to the system through the manual startup. The second process considers *all* edges, and, using the correspondences found by the first process for the particular line correlation it is presently performing, suggests a larger set of correspondences. Those correspondences which are seen to 'persist' over several preceding second process line analyses (implying that they arise from true scene geometric discontinuities) are given for consideration to the first process for its *next* line analysis.

The correlation's metric is pixel intensity difference. The two processes both use a least squares minimization on these intensity differences to choose the optimal edge correspondences. Edges are used to bound the linear regions, or intervals, being correlated, and edge correspondence is a side effect of the intensity correlation — edges themselves are not compared.

The algorithm progresses from one image epipolar line to another, propagating results (to limit subsequent search) as it goes. The algorithm, as noted in the summary, requires manual starting. It propagates determined correspondences along paths of proximal edges as it progresses from line to line. Constraints have been built into the system to make it only applicable to planar surfaced structures, and the correlation only accepts transitions indicative of nearly horizontal or vertical walls ... in fact, they go to substantial effort to ignore surface detail (such as roads, sidewalks, windows). The algorithm preprocesses the imagery data in a way that precludes it from working with anything other than straight lines (as derived from sequences of edges) in the images. They have processed and documented the analysis of a single scene with their algorithm.

Their aim was to produce a three-dimensional planar rectilinear description of cultural scenes. The results shown do not indicate that they have succeeded. One point to note is that their use of two correspondence processes, with the second introducing 'new' and removing 'old' scene structures from the analysis, introduces a hysteresis into the processing — new structures (in the direction of processing) take a while to be believed ('persist'), while old structures take a while to disappear once passed. Precision would seem not to have been one of the desired properties of their system. Further, a recent paper from the group comments on the instability and 'noisy' nature of the two-process structure ([Degryse 1980]), and explains several constraints they propose introducing to reduce the effects of these problems (see also [Panton 1981]). The constraints — the scene is imaged orthographically, the structures are strictly rectilinear, all vertical surfaces are either parallel or orthogonal, and all horizontal surfaces are parallel — are severely restrictive, and have no provision for the generality and flexibility a reasonable stereo system must have. Once introduced into the analysis, it is difficult to conceive of how these restrictions could be removed for the processing of more general domains. The constraints they have used serve to bound the applicability of their process, rather than bounding its cost.

These criticisms aside, however, there is a lot of merit to their work: the overall approach they took was fairly comprehensive, and they addressed many important imaging and correspondence questions as side issues of their study. In the context of their goals, the constraints they introduced were reasonably valid; although one should note that the crucial question of identifying a scene as cultural in order to allow this constrained interpretation was not addressed. A benefit of having read the reports of this work was in noticing their use of dynamic programming for the optimization; a variation of this technique has made a considerable contribution to the efficiency of the correspondence process used in the research I will be discussing here (see [Forney 1973]). Dynamic programming for stereo correspondence was first documented in [Gimel'farb 1972].

### Marr-Poggio System 1980

The approach of the MIT group is in melding psychological theory and observations into a computational algorithm for stereo vision. They consider neurophysiological relevance and biological feasibility crucial aspects of their algorithm, and support the details of their approach with extensive references to the perceptual psychology literature. The algorithm, developed basically by David Marr and Tomaso Poggio [Marr 1977], is an edge-based line-oriented filtering and matching process. Grimson's implementation of the stereopsis algorithm [Grimson 1980] processes as follows:

- Fill 4 pairs of working arrays with zero-crossing values and orientations. The zero-crossings are found by convolving the images with 4 spatial frequency tuned band-pass filters, varying in size from 7 to 63 pixels in width.
- Set initial vergence values for the eyes in the two images (manually).
- Match zero-crossings in the paired arrays with these relative eye positions. Within paired arrays, the process decides upon acceptable matches on the basis of zero-crossing *contrast* (positive or negative) and very rough edge orientation estimates (quantized to 30 degrees, so slopes must be within approximately 60 degrees of eachother). Matches are of ***positive, negative,*** and ***zero*** disparity, relative to the vergence.
- Mark ambiguous or 'no-match' edges as such.
- Check unmatched points in ***regions***, and for those where this number is greater than 30%, delete all matches. Regions are defined with regard to some statistical measure to ensure that the size represents a reasonable local sample.
- On the basis of low frequency filter matchings, make various ***positive*** and ***negative*** vergence movements to bring unmatched high frequency edges into correspondence (high frequency edges come from the smallest filters), and iterate on the matching process.

A subsequent process interpolates a smooth surface to this derived edge-based disparity data, resulting in a full depth map. The assumption that allows interpolation to take place is that ***'no information is information,'*** *i.e.* that the lack of edge signal in a part of the scene indicates that there are no intensity discontinuities there, and thus likely no depth discontinuities. If the scene contains no occlusions then this assumption is valid; although, even allowing this, it is rather dismissive of useful intensity data which could provide information on subtle surface shape variations. What the assumption principally neglects is the difficulty presented by **unseen** intensity discontinuities ... those hidden by occluding contours. In his work, Grimson presumes that an intensity discontinuity separates image locales of equivalent disparity. Counter examples abound. Having this ***'no information is information'*** assumption, the interpolation scheme makes no distinction between surface boundary points (where there is depth discontinuity) and surface detail (where there is none) ... the former should be breakpoints for the interpolation, the latter knots. The resulting surface fitting smooths an ***'elastic plate'*** over the entire scene. Elegant as the interpolatory analysis may be, the only interesting solution to the problem of defining inter-edge surface shape would be one which considers the global context at each edge ***('Is there any indication that this is an occluding edge?')*** and, where possible, domain knowledge ***('Are their buildings in the scene? Does this seem to be the top of one?')***. That is, an interpolatory technique must be coupled with a scheme to distinguish knots from breakpoints.

The results published include the analysis of several random dot stereograms, each composed of 4x4 randomly positioned black and white squares, with the maximum vergence variation running from

2 to 6 dot widths. Other examples include a ground level building scene, a view from a Mars Viking vehicle, and a random dotted coffee jar.

Assessment of the algorithm is a bit difficult: it uses a fairly simple control structure with unsophisticated matching criteria, and its success from these mechanisms is quite remarkable. But questions arise. The approach lacks a mechanism for assessing **global consistency** in its correspondence results. It would seem from the discussion of the algorithm that the initial eye vergence plays an important role in determining the final set of correspondences. By accepting high frequency channel correspondences on a *local* basis the implementation precludes other vergence matchings which could be *globally* more satisfactory (it should be noted that lower spatial frequency is **not** synonymous with *globality* — see [Julesz 1976]). Notice also that the low-frequency to high-frequency control structure that is said to be as used here is shown in [Frisby 1977] to be inadequate as a model for human stereopsis. Using a maximum filter size that corresponds to the largest observed in foveal vision only (the implementation doesn't vary filter size with eccentricity, as the theory suggests), Grimson has excluded from his processing the possibility of the more globally-driven radical vergence movements that seem necessary for scenes having large disparity variations. Perhaps this would be recoverable through the correct implementation, with filter size varying with eccentricity ... he has only implemented the theory for foveal vision. **Monocular cues**, which their theory doesn't address, are known to provide information for such radical vergence movements ([Saye 1975]). Initial vergence is set manually; it is not clear how subsequent major vergence adjustments are controlled. In fact, several control strategies are experimented with in the text, each to give the optimal results for the channel noise settings being tested. No clear definition of vergence control is given. In the light of the chronic failure in past vision research to document limitations and test to the breaking point, it may seem rather unfair to bring criticism to an apparently successful algorithm such as this, but its completeness has yet to be demonstrated (an interesting recent extension of the Marr-Poggio theory of stereopsis that addresses some of these issues is described in [Mayhew 1981]).

Dissatisfaction with the Marr-Poggio theory, and its implementation by Grimson, centre around:

a) their failure to define precisely its vergence mechanism,

b) the lack of a global control structure, one which would guarantee some optimal correspondence between the two images ([Frisby 1977]),

c) its failure to adequately consider other both local and global constraints in its matching criteria (such as statistical characteristics of surface slope, edge orientation, and intensity variation), and

d) the theory's neglect of established monocular cues to stereopsis ([Saye 1975]) – it would appear to be owing in large part to chance alone that images with large disparities could be fused correctly.

Although the approach to be discussed here isn't based on adherence to a theory of human stereopsis — rather, it centres on an analysis and exploitation of various geometric and photometric constraints on an imaged scene — parallels do exist between it and the Marr-Poggio algorithm. Both process *edge* descriptions of the image pair, determining correspondences on the basis of local edge properties, both work at several levels of image resolution (although with differing techniques), and both aim for a *depth map* description of the imaged scene.

### *Control Data Corporation Structural Syntax Approach 1981*

Two documents, [Degryse 1980] and [Panton 1981], describe more recent work from the CDC group that was designed to supplement their previous epipolar matching process [Henderson 1979], which they classified as 'noisy', 'fragmented', and 'unstable'. They hoped to introduce information of a more geometric nature to constrain the possible interpretations and "*remove some of the unreliability* *and ambiguity*" of the matching process. At the same time they redefined their goals so as to remove the urgency of the 'autonomous' in their processing. Again, they are concerned in this work with the analysis of images of urban structures, stereoscopicall; projected either orthographically or centrally to planar imaging surfaces. The [Degryse 1980] paper describes modifications to their Broken Segment Matching scheme, while the [Panton 1981] report describes subsequent work.

Noting the inadequacies of their first stereo system for processing in the same domain [Henderson 1979], calling it 'blind' to the surrounding context of the cultural scene, they argued that they needed to incorporate *a priori* knowledge of cultural scenes into their analysis. They designed a 'structural syntax' to provide this geometric information. The structural syntax is introduced as a set of geometric principles specific to the sort of 3-D cultural scene their research addressed. Intended application was restricted to structures in the form of right parallelepipeds; the structural syntax defined a mechanism for the restrictive interpretation of scenes as these objects.

There are three principal elements of their structural syntax, and these are shared by both recent approaches:

1) The edge orientation principle uses the convergence of 3-space parallel lines to vanishing points for clustering parallel edges. The authors presume that building orientations are known and are all identical, so that a single pair of vanishing points suffices for all scene horizontal edges, and there is a single vertical edge vanishing point. In the [Degryse 1980] work, this labels edges, so limits the set of possible edge matches. Note that the syntax is being used here to restrict the projective orientation and shape of all scene surfaces. Vanishing points are currently determined manually (utilization of vanishing points for polyhedral scene interpretation has also been suggested in [Liebes 1981]).

2) The principle of known or fixed transform slope governs the allowable 3-space orientations of building faces, constraining surfaces to be either vertical or horizontal. This constrains the solution paths in their Broken Segment Matching process in [Degryse 1980].

3) The min-max transform principle limits the range of acceptable heights for structures to some interval known *a priori*, and is used in both the Broken Segment Matching process.

[Degryse 1980] showed no computed results. Testing of the algorithm specified in [Panton 1981] was done on a small portion of a single pair of images of one medium sized building.

The authors acknowledge that their systems still require extensive testing and development. The present systems appear to demand a substantial amount of skilled operator intervention, requiring iterative tuning of parameters and repeated passes through the low-level processes. As an aid to manual reference preparation either of these systems may be adequate. But neither will suffice where automatic and flexible processing is needed. As an example, note that the restrictions imposed by the 'syntactic rules', the need for manual intervention at almost all stages of the processing, and the lack of success at even this simply structured problem make these systems completely inappropriate for the real-time processing needed of the system that is to *use* the models created by such a 'reference preparation' system.

## 2.2 Critique of Existing Systems

### 2.2.1 – Autonomous processing

A stereo system to operate for autonomous mapping, reconnaissance, or inspection in some domain must be able to initialize itself and run without the need of operator intervention.

Of the systems described above, only Gennery's runs entirely autonomously. The system of [Panton 1978] appears to require manual initialization, as does certainly the Control Data Corporation systems [Henderson 1979, Degryse 1980, Parton 1981] and, to a lesser extent, the [Grimson 1980] system. These may also require manual intervention during the processing — the [Henderson 1979] and [Degryse 1980] when there are vertical breaks in scene continuity, the [Grimson 1980] when the disparity differences exceed the size of the largest convolution operator, and the [Panton 1978] system when the terrain approaches discontinuity and the correlator begins to diverge locally from the correct matchings.

### 2.2.2 – Domain restrictions

An understanding of its domain of intended use and an analysis of its performance capabilities will give us insight into a stereo system's overall range of application, and thus its utility.

In general, the performance of the area-based correspondence schemes will degrade rapidly when confronted with scenes of discontinuous structure, and this makes them inappropriate for the analysis of cultural sites. The CDC techniques of [Henderson 1979], [Degryse 1980], and [Panton 1981] exclude the processing of rolling, curved, or even non-rectilinear structures — predisposed to the analysis of building tops, they are inappropriate for most everything else. None of the systems described can work well where details in the background have reversed positioning with respect to occluding surfaces lying before them (consider a finger at arms' length and the background beyond) — this is referred to as the *edge reversal* problem. The Grimson work is the only one which does not make explicit mention of excluding such positional reversals between the two imaging planes, although it probably does so in the working of its region disparity consensus and its use of disparity pooling in the matching process. Excluding edge reversals is such a convenient expedient when working with epipolar geometries that it has been widely accepted for the correspondence processings. That it is a restriction becomes obvious when it is noticed that it prohibits the simultaneous fusion of a thin object (like a pole) and its background — relative to the pole, what is left-right in one image will be right-left in the other. This artefact of the processing may be excused to some extent in that it is also observed in human stereopsis, but there is no obvious necessity of building limitations of the human system into a machine system (in their study of the limitations of binocular fusion, Burt and Julesz ([Burt 1980]) comment on inability to attain fusion with positionally reversed points).

Looking at the range of examples presented in the published results from these stereo systems also provides insight to their applicability. [Panton 1978] has demonstrated a single rolling terrain stereo pair analysis, as has [Gennery 1980], although Gennery's scene contains some rather large rocks and the scene slopes off to a (not seen) horizon. [Levine 1973] shows the processing of two rock-strewn scenes, similar to that of Gennery. The views in these area-based systems are, as expected, of terrain, and depth discontinuities are either not severe or ignored. [Grimson 1980] has applied his algorithm to considerably more scenes ... many random dot stereograms, and several real image pairs.

### 2.2.3 – Global consistency and monocular cues

The human perceptual system has the advantage that it can call upon higher processes to comment on the consistency of its visual observations. Only rarely is our binocular sight confused by ambiguities, and then this can usually be removed with a tilt of the head or slight motion to the side for a different perspective and more information (an observation which lead Moravec to his development of *slider stereo* [Moravec 1980]). An interpretation mechanism is at work with which our stereo systems at present have little to compare. Important considerations for a stereo system are how successful it is at resolving ambiguities, and how consistent is its interpretation over the entire scene.

Some researchers have decided that a *smooth* result is a good approximation to a *consistent* result, and perform local averaging of depth measurements, hoping to diminish the impact of gross errors through the abundance of good correspondences (for example [Levine 1973], and [Grimson 1980] with his disparity consensus requirement). A superior approach is to work within a set of valid assumptions or observations on the nature of the viewed world, and use the implications of these to choose among ambiguous or inconsistent interpretations. [Gennery 1980] uses an analysis of his correspondence error distributions to enable the automatic editing out of 'wild points'. One common assumption is that the scene is smooth and continuous most everywhere, and can be expected to be discontinuous at only a small number of locations (for example at those places where the viewed luminance is undergoing abrupt change).

The way such knowledge enters the analysis varies. In some work, the continuity assumption is used in prediction. [Levine 1973], [Panton 1978], and [Gennery 1980], in their area-based systems, use the context of neighboring points to limit the search for point correspondents, presuming that points neighboring in two dimensions should be neighboring in three dimensions. But this has determinacy problems - the results would change were the analysis to be done in a different order, for example with right to left scanning rather than left to right — and decisions are made locally, in a set direction, usually never to be revised. Further, these systems do not have mechanisms for locating actual scene depth discontinuities (see below). The MIT Grimson work makes good use of inference on the continuity of surfaces and the lack of edge signal in its interpolated surface fitting (see the summary), but again fails to deal adequately with actual scene depth discontinuities. Also, the system's use of context in its local edge matching is marginal, in that matching at a lower resolution (lower spatial frequency) appears to be a prerequisite[11]for matching at a level of finer detail (higher spatial frequency). A global metric is used in consistency checking of disparities over regions — requiring 70% of the disparities to be in agreement (one standard deviation, presumably), but this has been implemented without adequate analysis (see [Grimson 1980] page 75, where it appears to produce a highly quantized, planar effect). [Schumer 1979] discusses a possible mechanism in the human system for this spatial averaging of disparities.

### 2.2.4 – Identifying depth discontinuities

As suggested above, an issue related to the achievement of global consistency is the identification of depth discontinuities in the scene - those places where the viewed surface is not smooth and continuous. This capacity has not been reliably incorporated into area-based analyses, where poor matches arising from occlusions or extreme perspective effects merely return a low correlation value, indistinguishable from other causes of poor matches. In cases of occlusions, the intensity values in a window about the depth discontinuity in the two views would have little likelihood of corresponding, and here, the correlation coefficient as a measure of similarity is inappropriate. Edge-based analyses operate with the artefacts of (among other things) depth discontinuities, and the inference capability

[11] this may not be always the case; the system description does not make a precise definition of the control structure

is available here for distinguishing occlusions and abrupt changes in depth (although none of the cited systems use it). [Binford 1981] discusses the inference of spatial events from monocular cues.

### *2.2.5 – Parallelism possible*

A stereo system to be used for tasks of navigation or process control must be judged on its ability to provide depth measurements at rates approaching real-time. The enormous amount of computation inherent in the analyses makes it unlikely that a scheme with intrinsic ordered dependence in its processing will be able to provide adequate speed, The potential for parallelism in the algorithm is an important consideration.

Neither the [Panton 1978] nor the [Gennery 1980] approaches could take full advantage of the high parallelism possible in the computation since they process from left to right in columns across the match image, relying upon previous correspondences to constrain the search for matches. The [Levine 1973] and [Henderson 1979] approaches are similarly limited, in that they process by lines from image bottom to top, with each line progression passing up the results of the preceding line analyses to constrain the search. The Grimson algorithm is amenable to parallel implementation, and is in the process of being put into hardware ([Nishihara 1981]).

### *2.2.6 – Four criteria*

We would like a stereo mapping system to have:

- **no** necessity for manual intervention, either initially or during the processing,
- **no** domain bias — certainly no predilection to horizontal or vertical surfaces, and no limitation to strictly rectilinear structures,
- both *local* and *global* metrics, to enable optimization and confidence measures at both levels,
- a capability of being implemented in parallel hardware, with, for example, a simple partitioning of $n$ processors for $n$ lines of analysis, or a distributed array of $m \times n$ processors for a pair of $m \times n$ image arrays.

## *2.3 Goals of this Research*

As may be inferred from the critiques above, my intention when beginning this research was to design and implement an autonomous ***robust, domain independent*** stereo vision algorithm — one with a structure that would lend itself to a ***parallel realization***. These various aims were meant to be achieved in the following ways:

[*Robustness*] The information in a two-dimensional grey scale image is spatially highly redundant. Exploiting this, line-by-line processing would be used to obtain locally good correspondence estimates, and global consensus would be reached through a cooperative process that enforces three-dimensional continuity.

A two-dimensional grey scale image can be expected to have a broad spatial frequency spectrum. Filtering this spectrum and processing from the bands of lower frequency to higher frequency (in the direction of lower to higher noise sensitivity), provides the benefits of a *coarse-to-fine* control strategy ([Kelley 1970]). This suggests an analysis at several ievels of resolution, guiding the higher resolution matching from the lower resolution analysis. The hierarchic principle in this is intrinsic to the system's processing in several ways. Resolution variation is one of them. The general theme is to process first the most reliable signal and use this to guide the successively more noise-sensitive analyses.

[*Domain Independence*] The choice of general constraints (on general observations) as opposed to specifics of certain configurations, is the principal determinant of domain flexibility.

There are no assumptions on the nature of the viewed scene, other than that its structure doesn't vary between left and right imagings.

Testing of the algorithm on images of both cultural scenes and natural terrain would demonstrate this flexibility.

[*Parallel Implementable*] The algorithm should be designed so that its computational structure is partitionable into parallel streams. Local interactions only (in both the line-by-line matching and any subsequent global consistency process) would provide for the separation of computation along line-pairs. With such a structure, a machine with $n$ processors could be made to do $n$ lines of image analysis in time dependent only on line width.

The results of the processing should be a digital ***depth map*** of the viewed scene. This would produce three-dimensional data in a form appropriate for input to a three-dimensional terrain and/or object modelling system (such as ACRONYM [Brooks 1981a]).

These aims were all part of the initial design of the system, and have all been addressed in the research to be described here.

## 2.4 Summary of the Processing

The input to this system is two images forming a collinearized stereo pair. The collinearization is essential at present in that it guarantees that image lines correspond to epipolar lines (see [Hallert 1960]) — a constraint that greatly facilitates the matching process. The processing is begun by sampling the images in both horizontal and vertical directions, measuring the distributions of intensity values and first difference in intensity values. Intensity distributions are used to adjust image gain and bias, and the distribution of first difference in intensity is used to determine the intensity variance $\sigma_d^2$, a measure of image noise which has an important role in the correspondence process. Figure 2-1 shows a stereo pair of synthetic urban imagery provided by the Control Data Corporation. This stereo pair, as all pairs in this report, is positioned for cross-eyed stereo viewing.

In the first phase of its processing the analysis here is ***edge-based***. Edges are powerful abstractions of image content, and their use greatly reduces the combinatorics of the correspondence process. They provide higher precision disparity measures than intensity matching techniques, and, through their mutual connectivity, enable explicit use of global information for reducing the ambiguity at the matching level.

To obtain these edges the images are convolved with several operators to produce descriptions of the image intensity boundaries (edges) at several levels of resolution. The convolution operators work on a line of the image at a time, and consist of up to four zero-crossing filters and a low-pass smoother. The smoother is used to reduce the resolution of the lines of the images, halving resolution at each application. Such an approach has had previous successful application in visual processing (e.g. [Kelley 1970], [Marr 1977], [Moravec 1980], [Grimson 1980]), and has relevant ties to the neurophysiology of vision, where some researchers feel a multiple spatial frequency analysis is part of the human system's processing ([Wilson 1978a]) (although the filtering used here is low pass, and not bandpass). Reductions in image resolution are made until the image noise (as measured by the pixel intensity variance statistic) is less than one intensity unit. This resolution diminishing can proceed to a maximum of 3 reductions, at which point it has been found that, for the image sizes used, there is too little image content left to allow for reasonable matching. The filters detect zero-crossings in the second differences measured at each image pixel. Certain properties are associated with the edges found by these zero-crossing filters, and links are kept connecting the edges with those near them in the two-dimensional image. Figure 2-2 shows the full resolution edges found in the images of Figure 2-1.

Image lines are paired, corresponding ones from the left and right, and the edges contained in these are matched via a dynamic programming technique The correspondence process starts with the lowest resolution edges, and uses the disparities determined there to select which subsets of the full resolution edges will be brought together for possible matching. This mapping of low resolution correspondences to full resolution edges passes through the intermediate resolution depictions, although there are no explicit intermediate resolution matchings. Each pair of corresponding lines is processed independently. Figure 2-3 shows a typical pair of corresponding image lines, taken from the images of Figure 2-1.

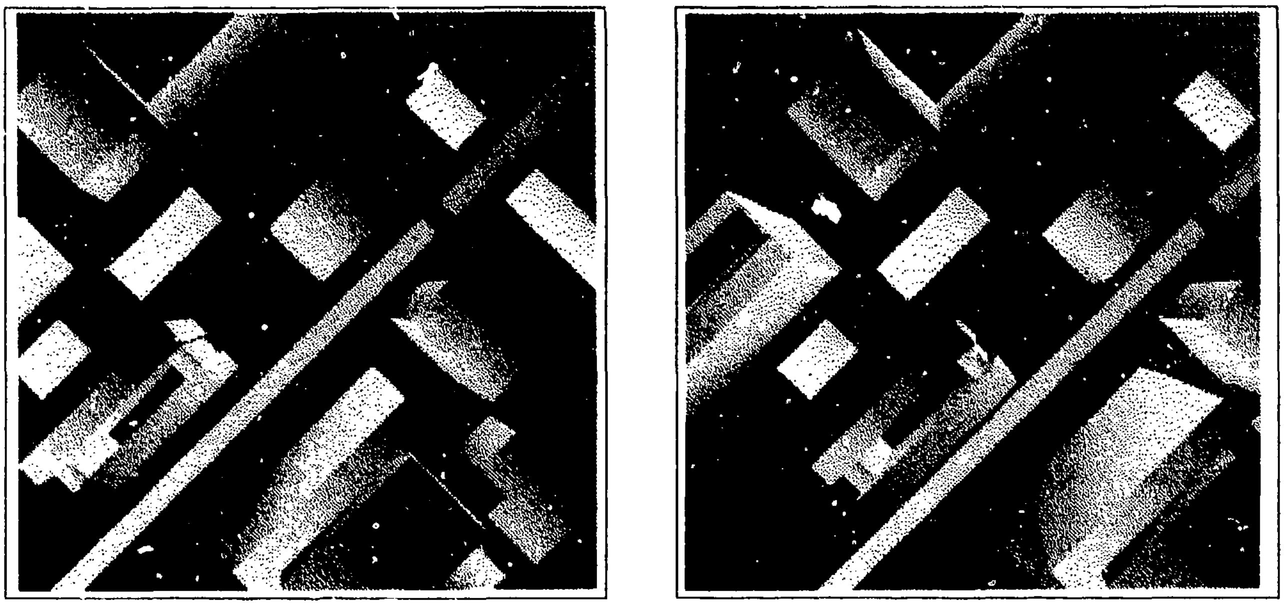
Once all lines have been processed and the various edge correspondences have been determined, measures of interline disparity variance are computed. Statistics of this are used to 'question' certain correspondences, and a cooperative process ensures that those inconsistent correspondences are removed.

This is the first half of the analysis, a low-to-high resolution matching of image edges with subsequent global consistency enforcement. It produces a reasonably dense edge-based disparity map of the viewed scene which forms a *template of constraints* for a subsequent correspondence analysis. Figure 2-4 shows (in stereo) the connected edge correspondences resulting from the processing of the images of Figure 2-1 to this point.

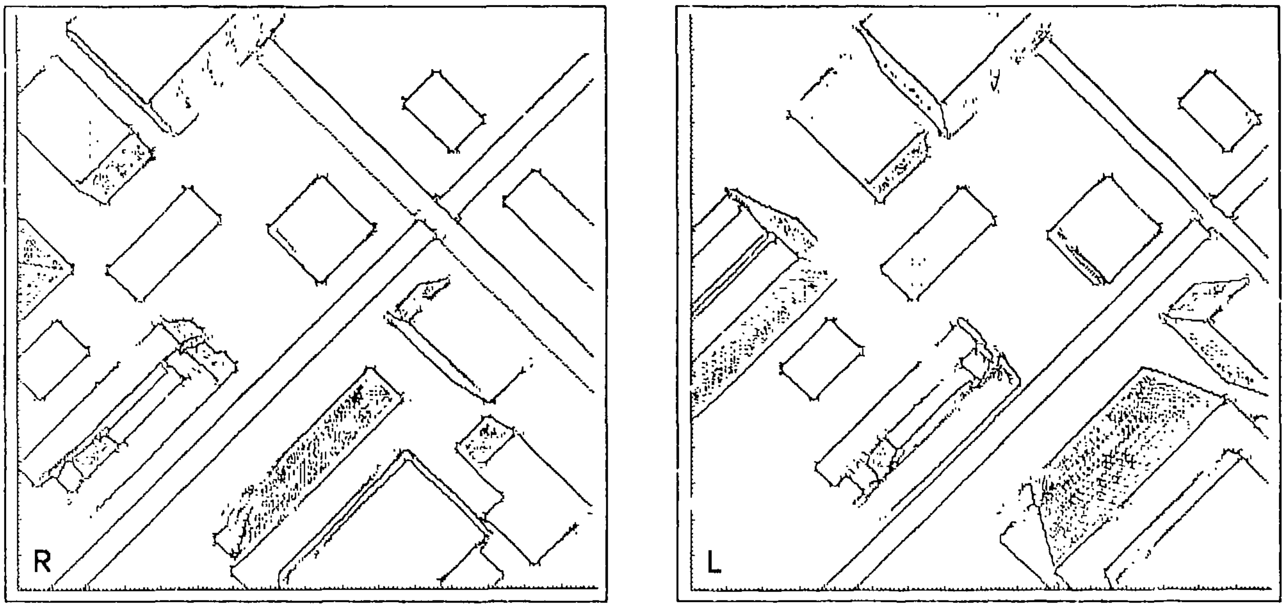
The second half of the analysis is a further **edge**, and then an **intensity**-based matching, and, as mentioned, these rely upon the first correspondence process's results to constrain the match possibilities. Selecting corresponding lines from the two images, the *edge-based* matcher attempts to pair edges which were either rejected by the earlier optimization process or were removed as 'questionable' during the cooperative consistency enforcement in the process of removing bad correspondences. Only those edges that are in corresponding intervals are considered for matching here. This edge matching completes the edge analysis.

The *intensity-based* matcher pairs not edges but image pixels themselves. It uses a metric which minimizes intensity variance and maximizes interpolated surface linearity. As in the edge-based correspondence process, the context of the matching is tightly constrained — corresponding pixels must come from corresponding intervals, as delimited by edge pairings. Intensity-based matching in general (for example [Hannah 1974], [Panton 1978]) is limited to analysis of rolling, smoothly varying terrain — it fails at surface discontinuities. *Edge-based* matching functions expressly at image locations experiencing high intensity variance, *notably at surface discontinuities.* So with *edge-based* matching providing precision disparity positioning and a highly constraining local context, the conditions are right for an *intensity-based* matching in the intervening intervals. Figure 2-5 shows the final elevation results of this processing for the images of Figure 2-1.

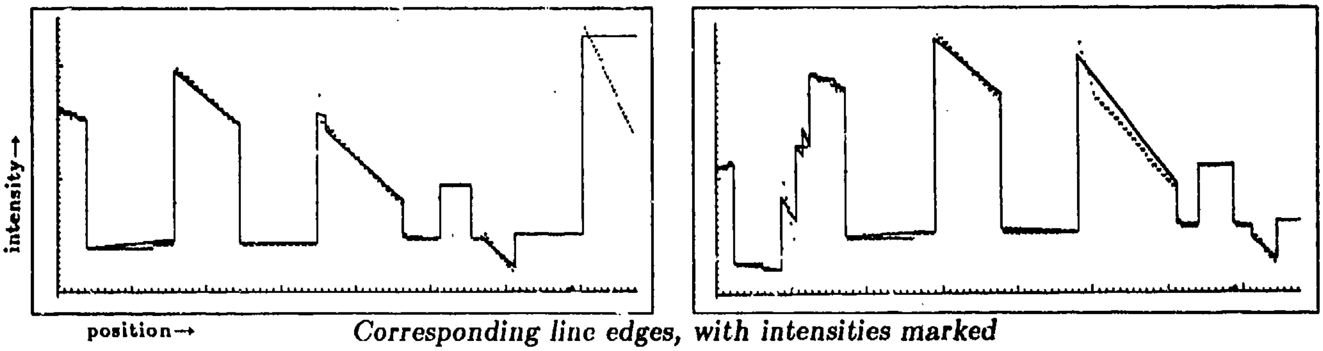
The matching algorithm in these last two cases is again a dynamic programming technique. The result of the full processing is a complete image array perspective disparity map of the viewed scene. Figure 2-6 highlights the structure and processing flow of this total scheme. A brief summary of the system can be found in [Baker 1981a].



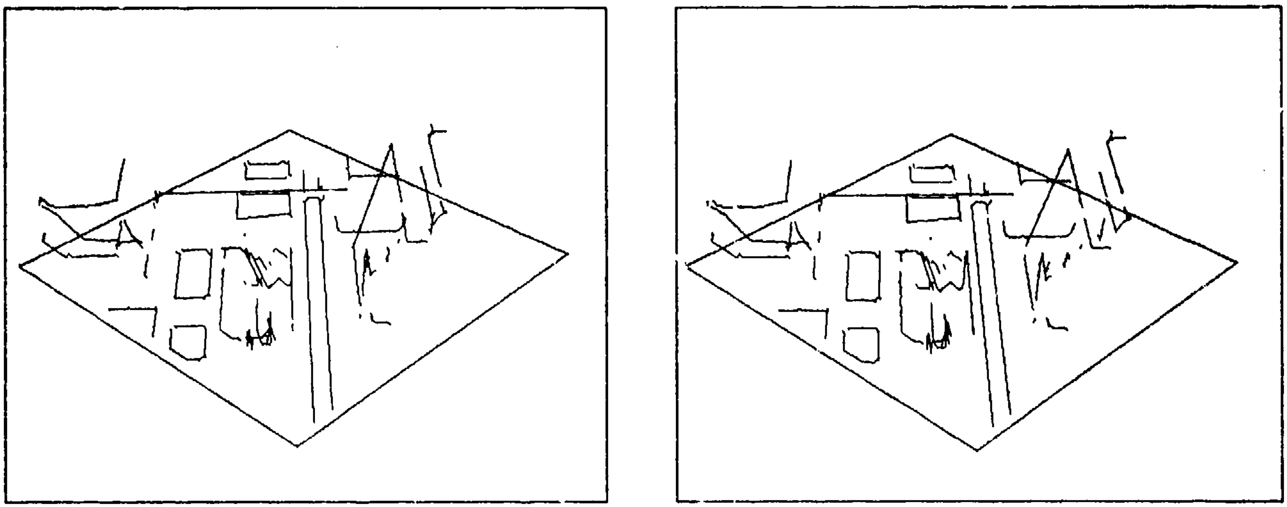
*A stereo pair of images (from Control Data Corporation) [256 × 256 × 6] (enhanced)*
Figure 2-1



*Edges of the stereo pair*
Figure 2-2



*Corresponding line edges, with intensities marked*
Figure 2-3



*Perspective view of connected edge elements*
Figure 2-4

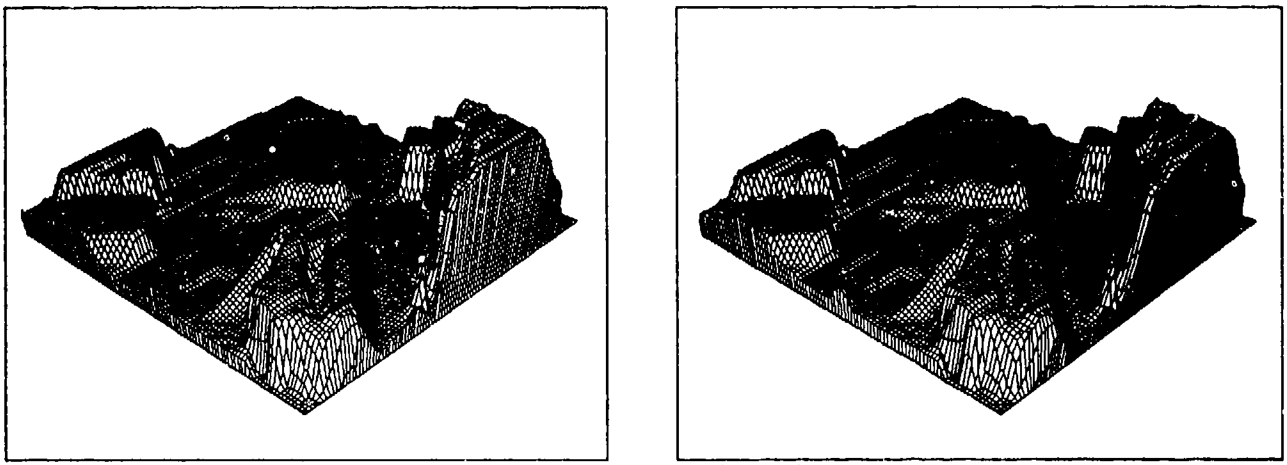
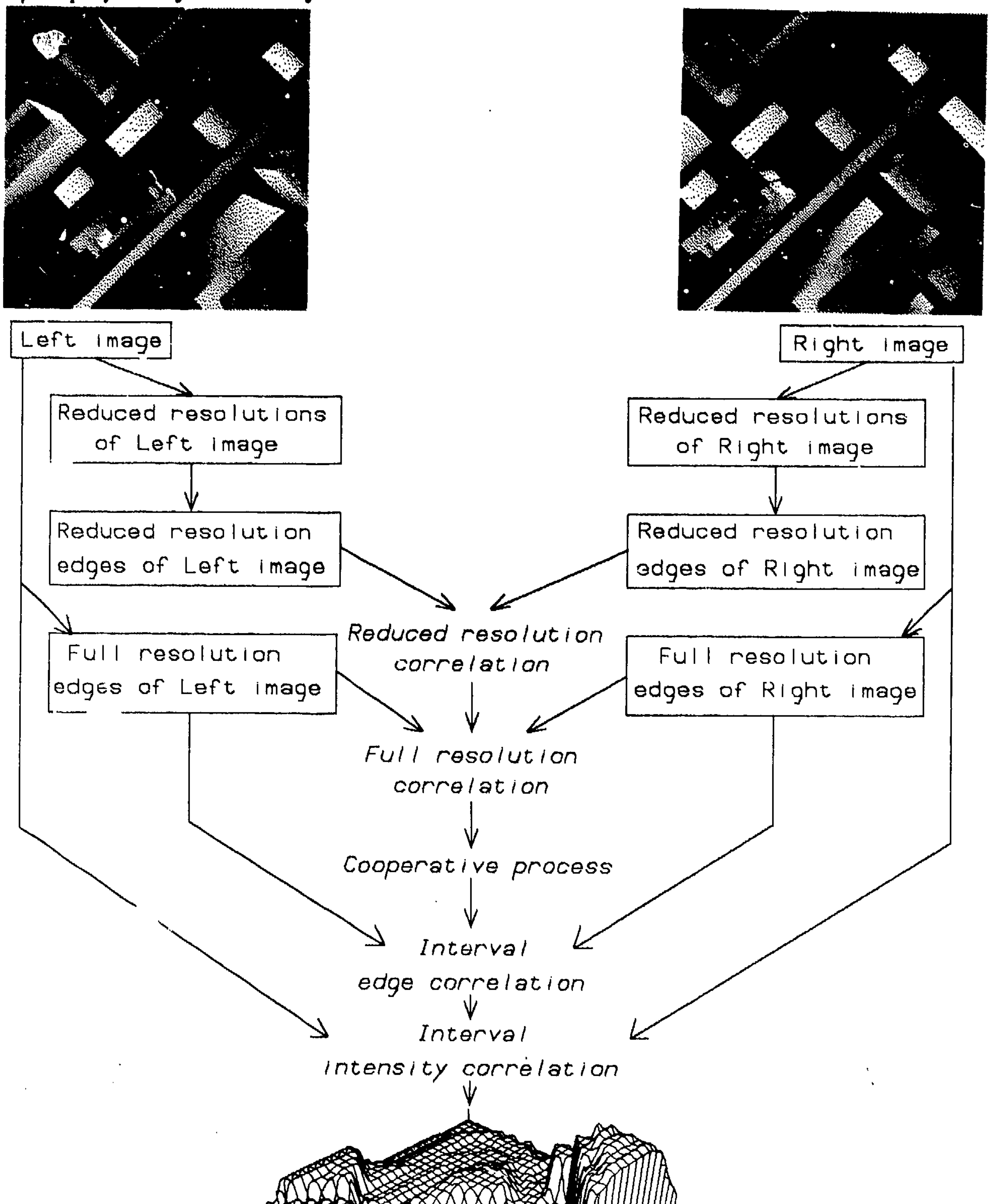


*Image array Orthogonal depth map*
Figure 2-5


Left image
Right image
Reduced resolutions of Left image
Reduced resolutions of Right image
Reduced resolution edges of Left image
Reduced resolution edges of Right image
Reduced resolution correlation
Full resolution edges of Left image
Full resolution edges of Right image
Full resolution correlation
Cooperative process
Interval edge correlation
Interval intensity correlation

System structure and image processing paths
Figure 2-6

# *Chapter 3*

# EDGES AND CONSTRAINTS

## *3.1 The Use of Edges*

The "edge-based" in the title of this report refers to the distinction between th. use of operators to reduce an image to a depiction of its intensity boundaries, which are then put into correspondence, and the use of area windowing mechanisms to measure local statistical properties of the intensities, which can then be correlated. The system described here deals with the former because of its:

a) reduced combinatorics — there are fewer edges than pixels,

b) greater accuracy — edges can be positioned to sub-pixel precision, while area positioning precision is inversely proportional to window size, and considerably poorer, and

c) more realistic invariance assumptions — area-based analysis presupposes that the *photometric* properties of a scene are invariant to viewing position, while edge-based analysis works with the assumption that it is the *geometric* properties that are invariant to viewing position). Edges are an abstraction of the image, are less sensitive to absolute image brightness levels, and highlight the structural aspects of the scene.

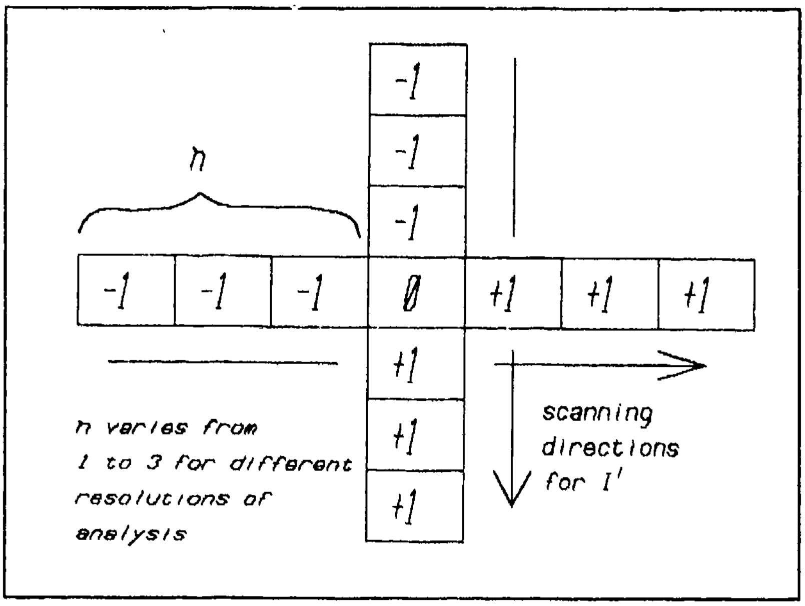
Edges are found by a convolution operator. They are located at positions in the image where a change in sign of second difference in intensity occurs. A particular operator, the one employed here for the full resolution analysis being 1 by 7 pixels in size[12](see Figure 3-1), measures the directional first difference in intensity at each pixel. Second differences are computed from these, and changes in sign of these second differences are used to interpolate zero-crossings (*i.e.* peaks in first difference). Certain local properties other than *position* are measured and associated with each edge — *contrast*, *orientation*, and *intensity to either side* — and *links* are kept to nearest neighbours above, below, and to the sides. It is these properties that define an edge and provide the basis for the matching. Correspondence techniques using similar such edge properties are described in [Marr 1976], [Arnold 1978], [Baker 1980], and [Mayhew 1981].

The operator processes left to right (horizontally) and top to bottom (vertically) in two separate passes over the image arrays, looking in each pass for oriented zero-crossings above a (noise-based) threshold (see Chapter 4, discussing statistical measures used in the analysis). Edge orientation is determined for each supra-threshold zero-crossing by the ratio of orthogonal components of the first difference operator, as shown in Figure 3-1. The left to right scan uses the horizontal component of this operator (7 × 1) and the top to bottom scan uses the vertical component (1 × 7).

## *3.2 The Use of Geometric Constraints*

The stereo matching is a search for **edge** correspondence between images. Figure 3-3 shows the edges found in the two images of Figure 3-2 with the second difference operator. The operator works in both horizontal and vertical directions, but it only allows matching on edges whose horizontal gradient lies above the noise — one standard deviation of the first difference in intensity. With no prior knowledge of the viewing situation, one could have any edge in one image matching any

[12] The edge operator is simple, basically one dimensional, and is noteworthy only in that it is fast and fairly effective.



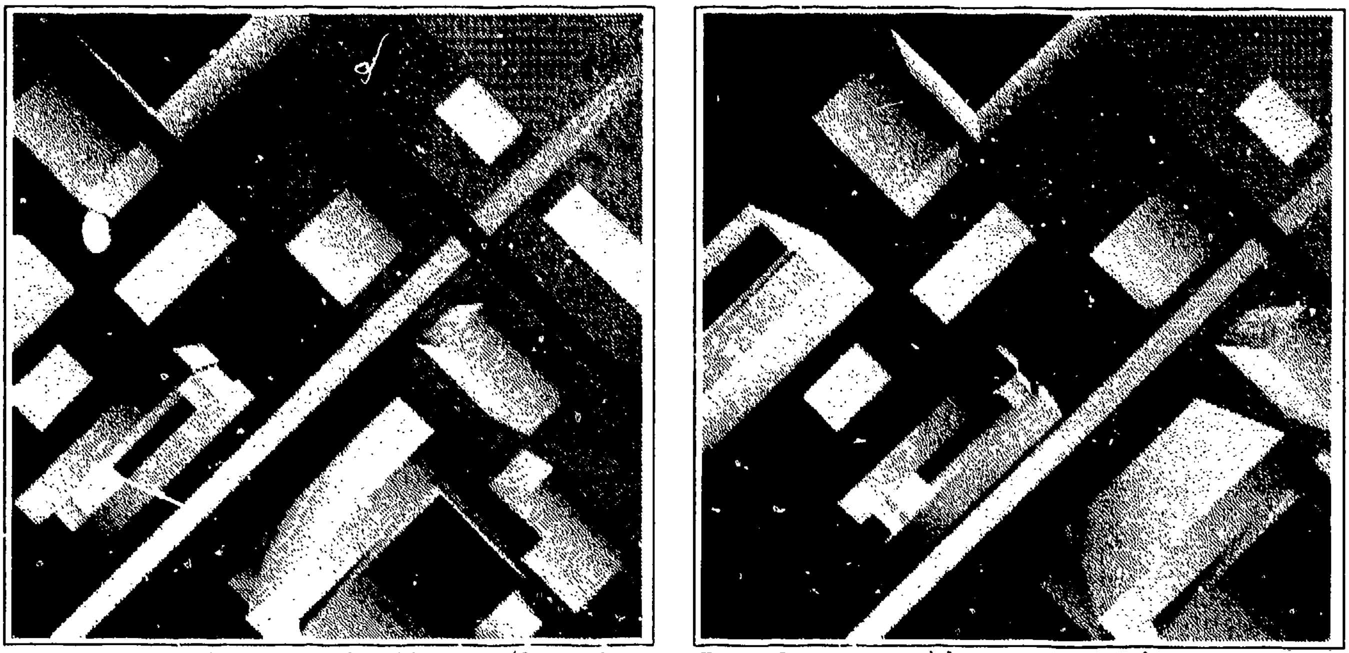
*Edge operator*
**Figure 3-1**

edge in the other. The combinatorics of this can, understandably, get very high. One would like to introduce general constraints to limit the cost of this search.

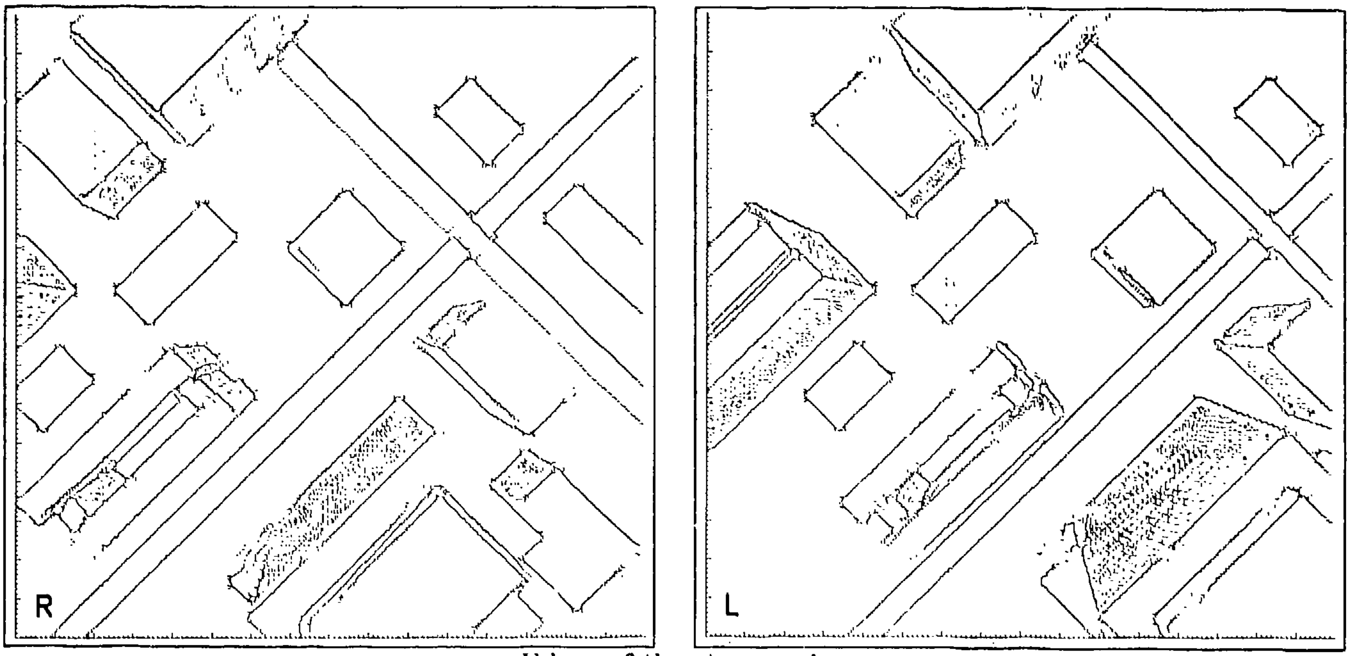
### *3.2.1 – Search constraints*

Knowing the geometric relationships between the cameras used in the imaging can greatly reduce the search needed in finding edge correspondences. Projective lines, termed epipolar rays, can be determined in the two images along which corresponding edges must lie. Figure 3-4 shows the geometry of this situation. With image planes $\pi_l$ and $\pi_r$ having principal points $P_l$ and $P_r$, imaging centres $C_l$ and $C_r$, line $C_lC_r$ is the ***epipolar axis*** through which pass all ***epipolar planes.*** The intersection of each epipolar plane with the two image planes $\pi_l$ and $\pi_r$ defines corresponding ***epipolar lines.*** A specialization of this general camera geometry is used, wherein the image principal horizon lines are collinear and the image principal vertical lines are parallel. In this configuration the epipolar axis does not intersect the image planes, and corresponding image horizontal lines are in fact epipolar lines. Although excessively restrictive for a general system, this was felt to be a justifiable simplication for our research work.

Consider Figure 3-5, in which two cameras are arranged in this configuration. Any point in the scene will project to two points on their image planes — one through each of the two lens centers (notice that the image planes are coplanar). The connection of these two points will produce a line parallel to the baseline between the cameras, and in this cas· parallel to the image horizontal lines. Corresponding edges in the two images, then, must lie along the same line in the two image planes. This camera geometry gives rise to images with a *collinear* epipolar geometry. The algorithm described assumes the stereo pair to be in a collinear epipolar geometry, and if this is not the case then the appropriate transformation of one image relative to the other must be made before further processing is done. Note that a less restrictive solution would be to have the correspondence process informed of the camera geometries, and have it solve for the more general epipolar geometry of

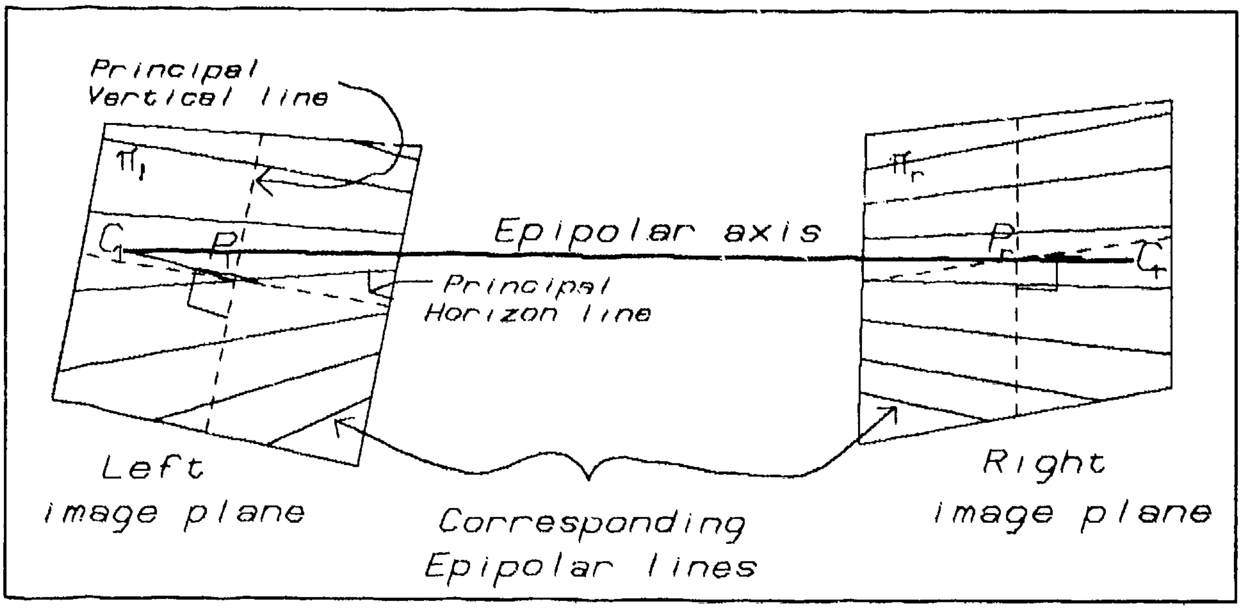


*A stereo pair of images (from Control Data Corporation)* [256 × 256 × 6]
Figure 3-2

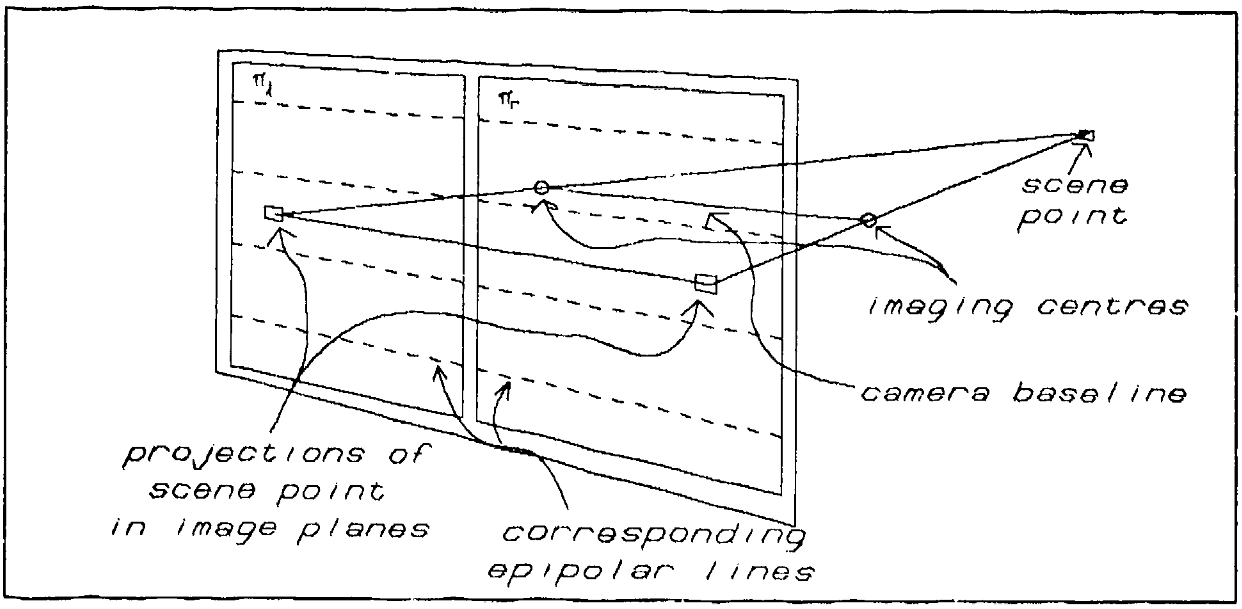


*Edges of the stereo pair*
Figure 3-3

Figure 3-4. Further refinements to this stereo process will include solving for the geometry at the matching level, rather than requiring it as a condition on the input.



General Epipolar imaging geometry
Figure 3-4

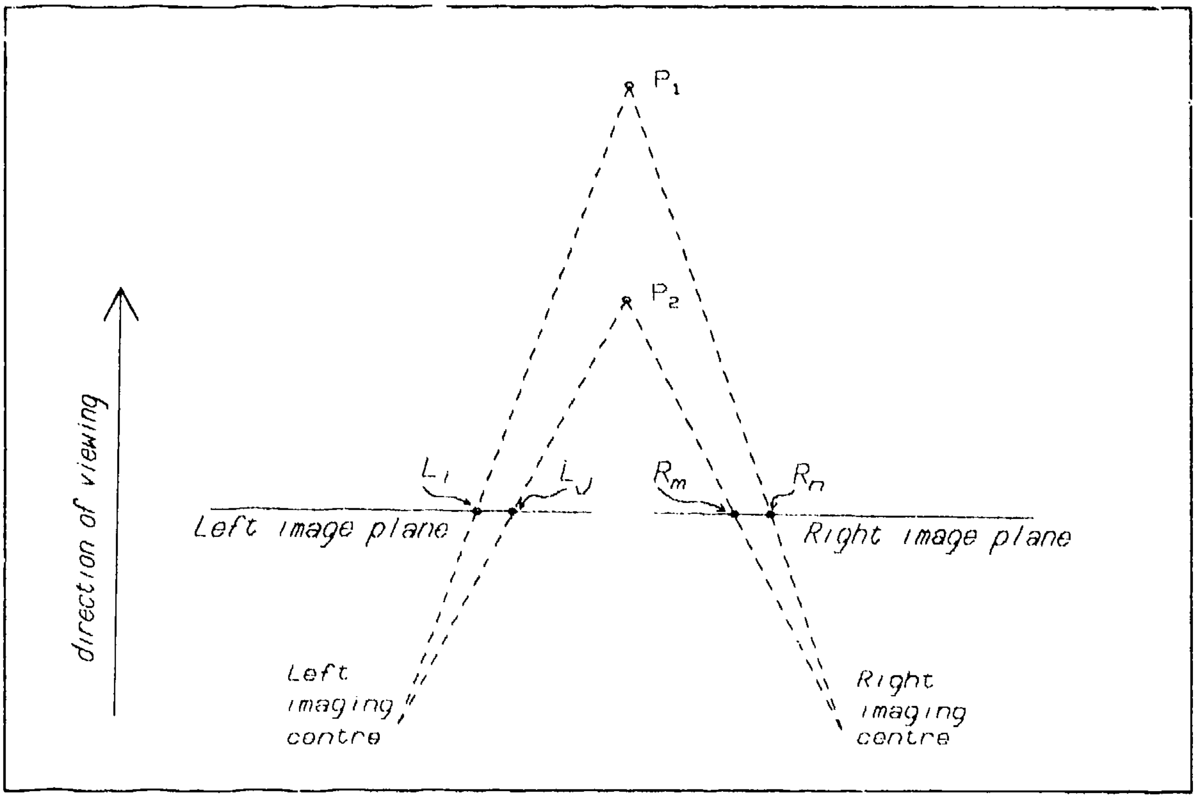

$\pi_l$
$\pi_r$
scene point
imaging centres
camera baseline
projections of scene point in image planes
corresponding epipolar lines

Collinear Epipolar geometry
Figure 3-5

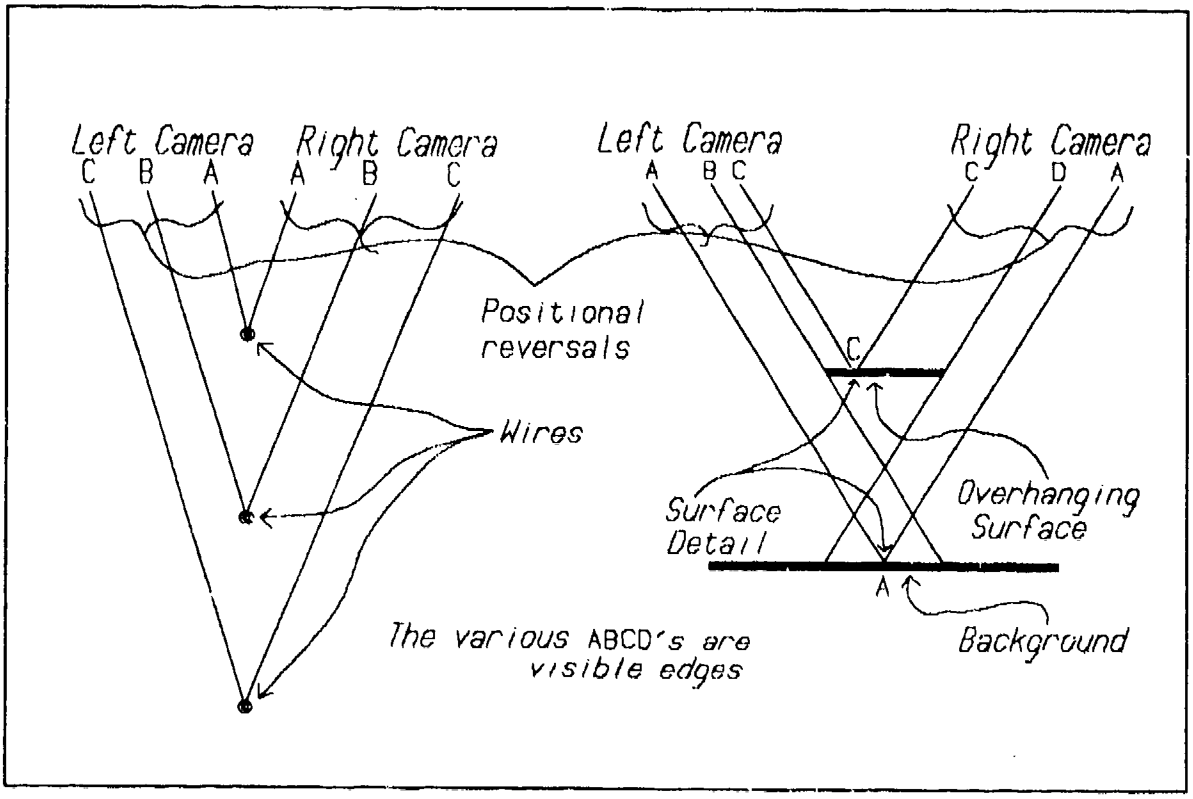
This discussion of camera geometry constraints suggests another crucial geometric constraint on the analysis. The matching algorithm to be described here demands a monotonicity[13] of edge order along

[13] The basis of this monotonicity constraint is explained in chapter 5 which discusses the Viterbi correspondence algorithm.

epipolar lines. This means that there cannot be reversals of edge order from one image to the other. Consider Figure 3-6. Left image edge $L_i$ which lies to the left of edge $L_j$ cannot match right image edge $R_n$ if $R_n$ lies to the right of edge $R_m$ which matches edge $L_j$. This constraint lies at the heart of the Viterbi method, although it is not without its drawbacks. Notice that if the image planes $\pi_l$ and $\pi_r$ face eachother, then objects in one image will be sequenced from the left while those in the other will be sequenced from the right. If the edges of these objects were allowed to match, it would violate our monotonicity constraint. This is a degenerate example of a general problem. The ordering of objects in the two projected images depends upon their distance from the imaging points — foreground/background appear as right-left or left-right depending on the camera site, and it should be clear that the problem of edge reversals is unavoidable. The use of this constraint will exclude from analysis, for the time being, such features as *wires* or *overhanging surfaces*, features which lead to these positional reversals in the image (see Figure 3-7). Psychophysical evidence suggests that this reversal also causes the human vision system trouble — we can fuse one or the other, the nearer or the farther, but not both at the same time ([Burt 1980]). Fusion of the items causing the reversal can be achieved only by vergence movements executed explicitly to bring them one at a time into fixation. (A similar method would provide a means of dealing with reversals here — reprocess the edges left unpaired by the matching process, treating them as satellites possibly left unmatched because of such local rivalries.)



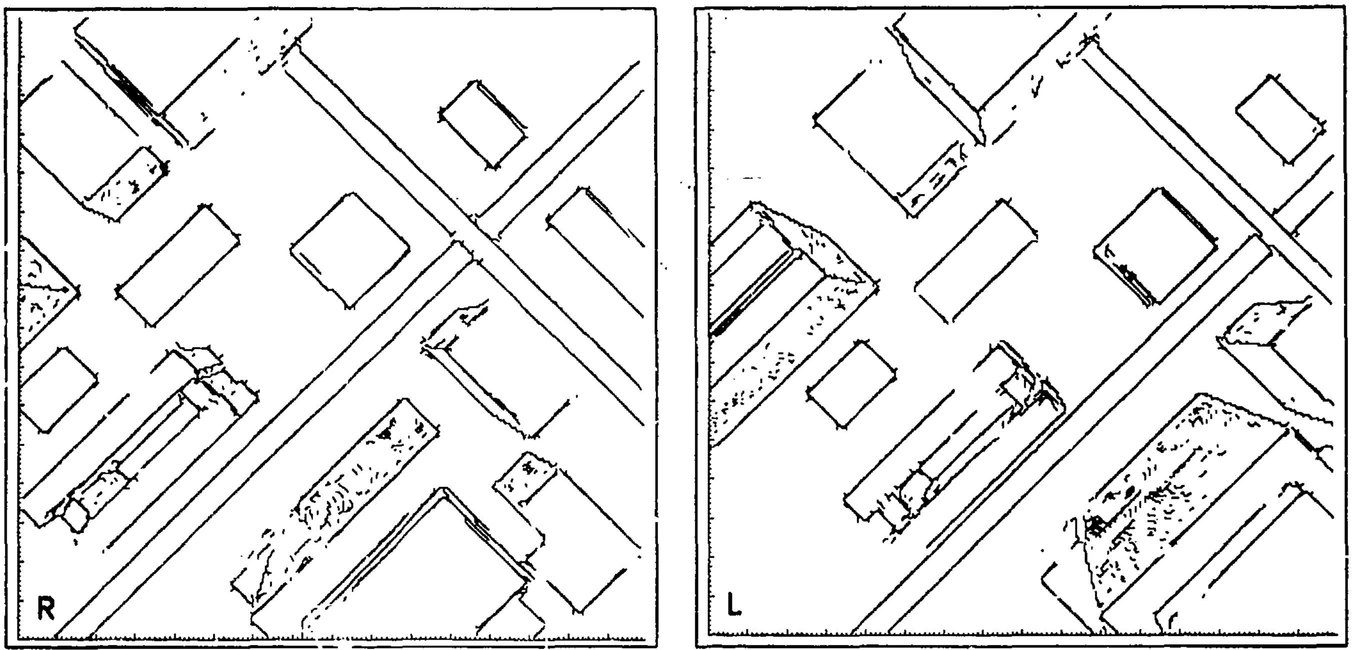
*Edge reversals along an epipolar line-pair*
Figure 3-6



*Examples of surfaces violating the Viterbi monotonicity constraint*
Figure 3-7

### *3.2.2 – Interpretation constraints*

When the edge-based correspondence has finished, it has come up with a judgement on which edges in the right image match a set of the edges from the left image. This determination is made on the basis of information strictly *local* to each line processed — there is no information made available to the matching from outside of the line to which it applies. Being so local, it has no guarantee of being *globally* correct, yet it is global correctness that we are trying to achieve. A very strong **global** constraint that can be of use here is that of *edge connectivity* (Figure 3-8 shows the connectivity of the edges of Figure 3-3). It may be presumed (by general position) that, in the absence of other information, a connected sequence of edges in one image should be seen as a connected sequence of edges in the other, and that the structure in the scene underlying these observations may be inferred to be a continuous surface detail or a continuous surface bounding contour. The individual line correlations make their suggestions of which edges correspond, and a subsequent cooperative process takes these local judgements and the known connectivity and works toward a global consensus. Statistics are kept (see section 4.2) on interline disparity differences along connected sequences of edges, and these measures, where a large disparity difference implies a large change in depth, provide the evidence for removing edge correspondences which violate observed bounding contour continuity.



*Edge cc nnectivity of the stereo pair*
Figure 3-8

### *3.2.3 – Constraint summary*

The three principal constraints on the analysis are that:

- the geometry of the cameras be known, and in particular, be the specialized geometry where image lines correspond to epipolar lines,
- there be no edge reversals along epipolar lines (if they are present, the solution will involve a monotonic subset of them).
- edge *correspondence* be consistence with edge *connectivity* in the images (as these suggest depth continuity).

# *Chapter 4*

# STATISTICS

## *4.1 Correspondence Statistics*

The best solution for a matching will be determined on the basis of some evaluation function. The evaluation function takes local quantitative measures of correspondence likelihoods and produces a global score for a potential solution. Statistical measures play a large role in determining these local quantitative measures. In the first case one wants to be able to distinguish edges or intensity variances that are in some sense *valid* from those that may be merely *spurious* or a product of the digitization or imaging processes. Further, one will want to compare edge parameters and intensity values across images, and have quantitative means for estimating their correspondence likelihoods. For these tasks, we need some measure of significance in intensity variation.

### *4.1.1 – Intensity variation*

A pixel's brightness is measured as the integral of a weight function (for example a Gaussian) over the local intensity surface. The principal variation, or noise, in a pixel's intensity arises from characteristics of the sensor used. This variation is referred to as sensor noise,[14] and it may be modelled as a Gaussian process whose statistics may be estimated by measuring the distribution of interpixel intensity differences. Say that the variance of interpixel differences — determined by sampling first differences in horizontal and vertical scanning directions — is $\sigma_d^2$, so that its standard deviation is $\sigma_d$ (zero-mean), then the variance in a single pixel's intensity value may be given as $\frac{\sigma_d^2}{2}$ and its standard deviation is

$$\sigma_i = \frac{\sigma_d}{\sqrt{2}}. \qquad (4-1)$$

This measure (standard deviation in pixel intensity variation) is used for several image dependent computations. The full resolution edge operator (having width $w = 2n+1, n = 3$) could be expected to have a standard deviation in its difference values of

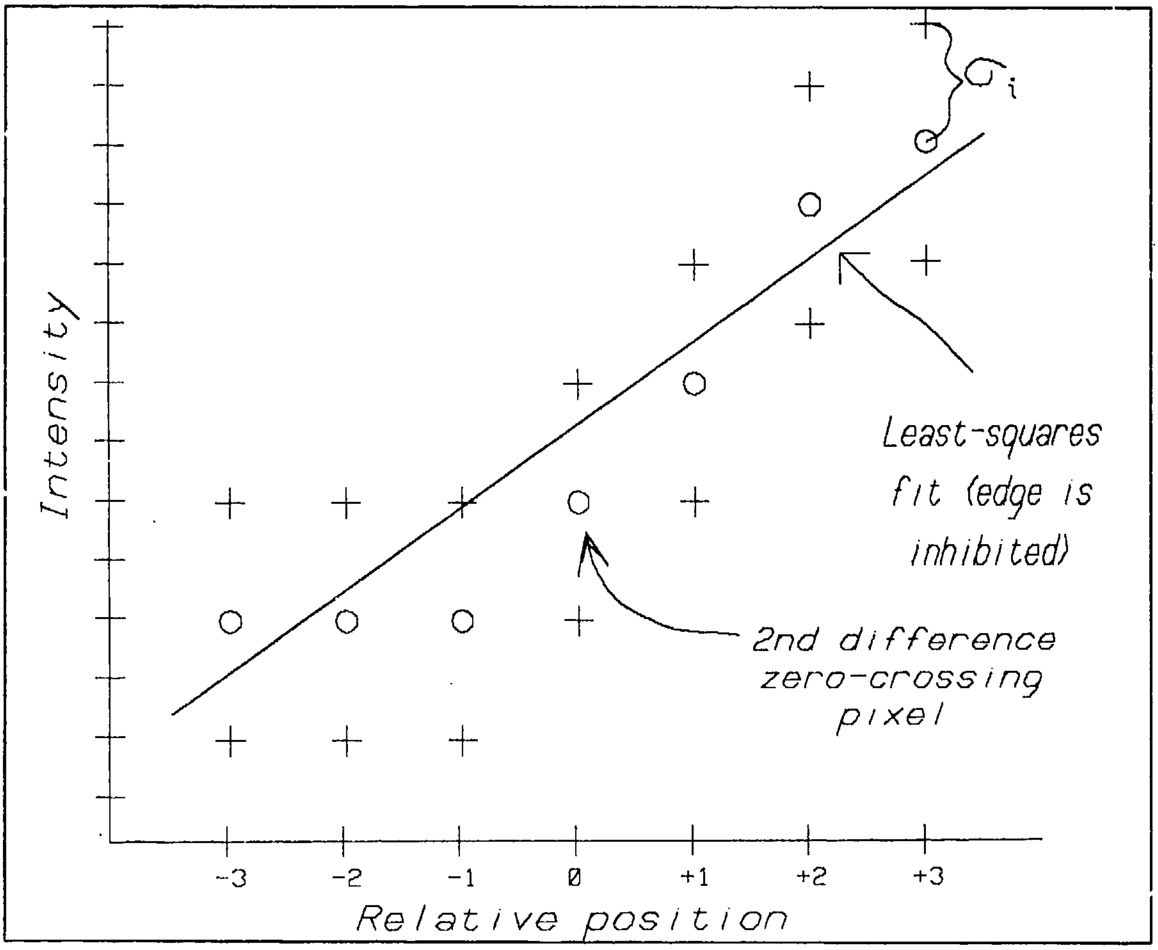
$$\sigma_f = \sqrt{2n}\sigma_i = \sqrt{6}\sigma_i. \qquad (4-2)$$

It is a zero-crossing operator, locating *edges* only at those pixels having a zero-crossing in their second difference (as defined earlier). However, discretization and camera noise make it necessary to look at more than just this zero-crossing measure. There can be areas where slight noise effects make the second difference fluctuate back and forth about zero, giving a great density of zero-crossings. A first difference threshold, based on the operator's intensity variance statistic $\sigma_f$, is used to separate *valid* edges from such noise-induced *spurious* edges — it ensures that the contrast across an edge is greater than $\sigma_f$, *i.e.* the matching will only deal with edges that are stronger than the noise. A further complication arises in that the signal variance measured here is not just a function of local image noise but of course of local image content as well. If the intensitv values are changing monotonically in some local area, as on a long gradual slope, discretization noise can give rise to zero-crossings in the second difference and the first difference measure $I'$ will, if the gradient is steep enough, exceed $\sigma_f$. A technique to remove local image gradient content is to apply a lateral inhibition operator to

[14] Shot noise, which varies with the signal, is not considered here, but if its characteristics are known then its noise effects can be compensated for by transforming the brightness values via a nonlinear function.

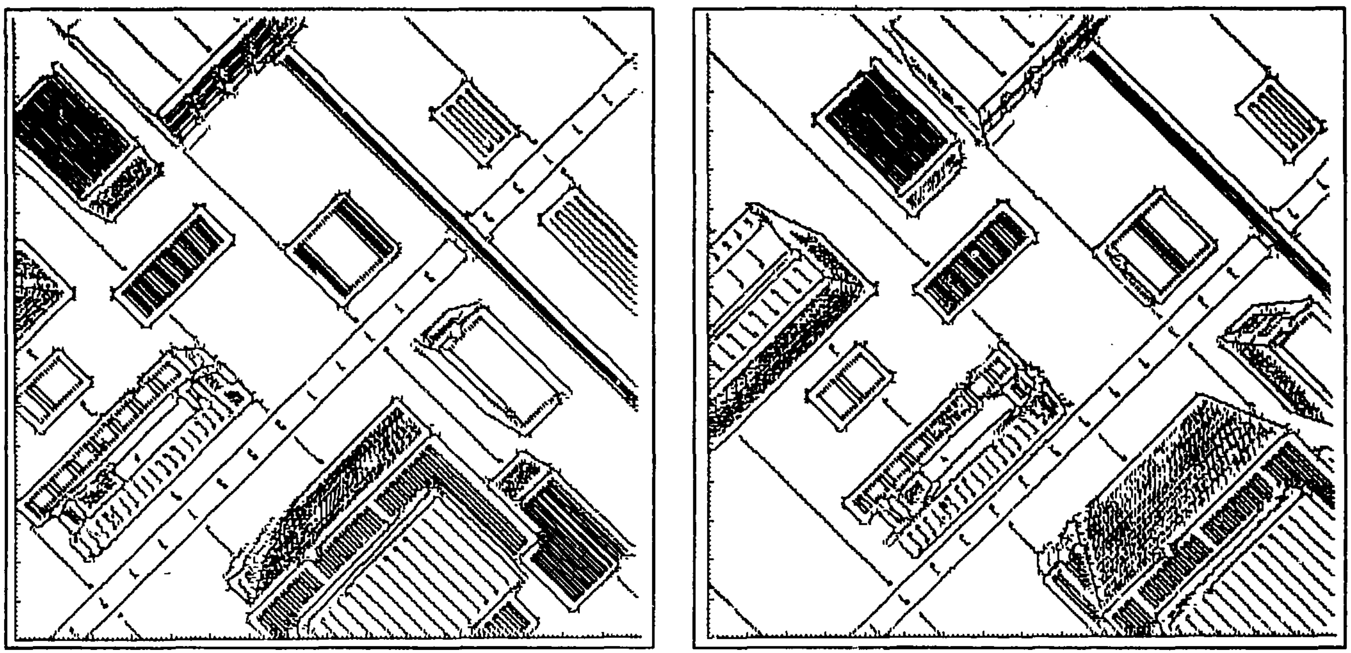
the signal ([Binford 1981]). This maps a linear function onto zero (*i.e.* it maps constant gradients onto zero).

A variation of this method is used here. At positions where there is found to be a zero-crossing in the second difference, a least squares line is fit through the support of the first difference operator. Each pixel intensity value $I$ and its standard deviation $\sigma_i$ defines a $[I - \sigma_i, I + \sigma_i]$ local error interval (see Figure 4-1). If the linear least squares fit to the intensity data passes through this $w$-length corridor, then the proposed edge on which the operator is centred is deemed to be *laterally inhibited*, and is not maintained as a valid edge. Figure 4-2 shows the output of the convolution with the lateral inhibition operation turned off — compare this with the edge set after lateral inhibition, as shown in Figure 3-3. This implementation of the lateral inhibition operation is basically an expedient, doesn't fit the normal mold of a lateral inhibition operator, and, in being only one dimensional, fails to take into consideration the more global structure of the image afforded a two dimensional operator. Its good characteristics are that it is evaluated only at candidate edges and, being centred on a symmetric operator, is very easy to compute.[15]



*Lateral inhibition operator*
Figure 4-1

[15] Further refinements to this stereo process should include giving both the lateral inhibition and the low-pass filters two-dimensional support.



*Preliminary edges of the stereo pair (before lateral inhibition)*
Figure 4-2

The standard deviation in pixel intensity, $\sigma_i$, can also be used to determine the accuracy of edge positioning. Recall that edge position is specified by an interpolation of the zero-crossings of image intensity second differences (see Figure 3-1). The standard deviation in the second difference measure is $\sigma_2 = \sqrt{4}\sigma_i = 2\sigma_i$. It is clear that the precision of the edge positioning depends upon this parameter $\sigma_2$ and the second difference contrast across the edge $C = |I_1'' - I_2''|$. Since the variation in intensity value is being modelled as Gaussian, we can determine the joint distribution of the variation in the quotient $x = \frac{I_1''}{C}$ as the convolution of the two normal (and equivalent) intensity second difference variation distributions with mean zero and standard deviation $\sigma_2$. Considering Figure 4-3, an error interval $[-\sigma_2, \sigma_2]$ can be defined about the interpolated edge position. The probability that the correct edge position passes within $\sigma_2$ of the interpolated position is

$$\int_{-\infty}^{\infty}\int_{-2(x+\sigma_2)}^{2(x-\sigma_2)} f(x)f(y)dxdy = 0.84166$$

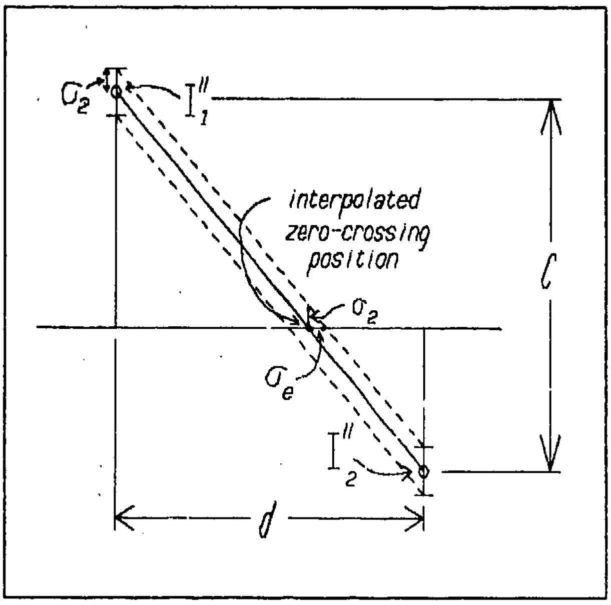
where $f$ is the Gaussian probability density function of (4-10) with $\eta = 0, \sigma = \sigma_2$. This is the integral of the convolution of the distributions in second difference variation, as Figure 4-4 may clarify. A convolution of Gaussians is Gaussian, so the variation of this convolution has standard deviation

$$\sigma_v = \frac{\sigma_2}{1.41} = 1.42\sigma_i.$$

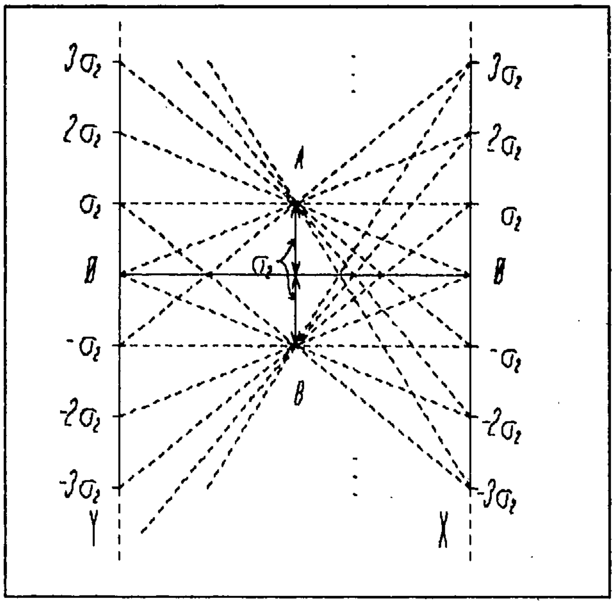
This is a measure of the vertical variance in interpolated position (as Figure 4-3); the horizontal variance in edge position can be determined from this as

$$\sigma_e \approx \frac{d\sigma_2}{C} = \frac{1.42\sigma_i}{C}. \qquad (4-3)$$

where $d$ is the distance between second differences, $d = 1.0$. This is a family of distributions with dependence on the measured pixel noise $\sigma_i$ and the contrast at the edge $C$.



*Interpolated edge position accuracy*
Figure 4-3



*Convolution of second difference variation*
$Y$ and $X$ are the variations in intensity second difference of $I''_1$ and $I''_2$, respectively, of Figure 4-3.
Figure 4-4

### *4.1.2 – Edge-based correspondences*

The edge-based correspondence process uses the pixel intensity variance $\sigma_d^2$ as one consideration in evaluating the probability of two edges corresponding. If the distribution of $I'$ values is Gaussian, then intensity differences can be mapped via the Gaussian cumulative distribution function to obtain a probability $\boldsymbol{P}_{ij}$ that left image edge element $E_{l,i}$ (which for brevity may be written $L_i$) with, say, intensity value $I_{l,i}$ corresponds to right image edge element $E_{r,j}$ (which may be written $R_j$) with intensity value $I_{r,j}$.[16] In the full resolution matching implemented here each edge $L_i$ is treated as **two** half-edges — the **left** side $EL_{l,i}$ and the **right** side $ER_{l,i}$ — and the intensity values $IL_{l,i}$ and $IR_{l,i}$ are the sums of the three pixel intensity values centred exactly 2.5 pixels to the left and the right, respectively (see Figure 3-1). This selection of intensity values removed from the edge functions to stabilize the metric, keeping those values in the area of high gradient nearest the edge out of the calculations.

The smoothing operator used is a 4 × 1 convolution with weights 1-2-2-1. When invoked to halve line resolution $t$ times it gives each pixel in the resultant depiction a support

$$S(t) = 3 \times 2^{t-1} + S(t-1), \text{ where } S(0) = 1,$$

[16] Of course an edge doesn't have an intensity; it has an intensity and a contrast, or two intensities.

in the original image. The standard deviation in intensity value for any pixel at resolution $t$ is

$$\sigma_i^T = \frac{\sigma_i}{\sqrt{2S(t)-2}}. \qquad (4-4)$$

The standard deviation in intensity difference at resolution $t$ for a first difference operator whose support is $2n+1$ is

$$\sigma_f^T = \sqrt{2n}\,\sigma_i^T = \sqrt{\frac{n}{S(t)-1}}\,\sigma_i \qquad (4-5)$$

(for the various smoothing operators used he'ε, $1 \le n \le 3$). These standard deviations can again be used to map intensity differences to c⁄rrespondence probability estimates via the Gaussian cumulative distribution function (they are zero-mean). The reduced resolution edge operators use these measures in separating valid from spurious edges, and the reduced resolution correspondence process uses them in estimating the likelihood of edges matching. *(Note: throughout, a superscript of T will distinguish parameters of reduced resolution t from those of the full resolution analysis.)*

These are the intensity statistics used in the edge finding and the correspondence processes. Other statistics are involved as well. The three edge-based matching schemes — *full resolution, reduced resolution*, and *constrained-interval* — have differing sets of statistically based metrics for measuring the likelihood of edges matching in their separate domains. In the following, probabilistic measures, parameters, or data structures are denoted by the prefix $P$, and the various multiplicative terms are independent.

### Reduced Resolution Correspondence

For reduced resolution matchings, at resolution $r = t$ with support $2n+1$, the probability that the edge $L_i^T$ corresponds to edge $R_j^T$ in the other image is estimated as:

$$PReduced_{ij}^T = PStat_{ij}^T \times PInterval_{i,p(i),j,p(j)}^T \qquad (4-6)$$

with $PStat_{ij}^T = PLeft_{ij}^T \times PRight_{ij}^T \times PContrast_{ij}^T$,

and $PInterval_{i,p(i),j,p(j)}^T$ is the probability that the interval between $L_i^T$ and its predecessor $L_{p(i)}^T$ corresponds to the interval between $R_j^T$ and its predecessor $R_{p(j)}^T$ in the other image (see Figure 4-5) — $p(i)$ is not meant to equal $i-1$, but rather is the predecessor along the path from $i$ that has a correlate $Correlate(p(i)) = p(j)$ in the other image,

where $PLeft_{ij}^T$ is the probability that the *left* intensity value of edge $L_i^T$ corresponds to the *left* intensity value of edge $R_j^T$ in the other image, and is computed as:

$$PLeft_{ij}^T = GPROB(EL_{l,i}^T - EL_{r,j}^T), \qquad (4-7)$$

$PRight_{ij}^T$ is the probability that the *right* intensity value of edge $L_i^T$ corresponds to the *right* intensity value of edge $R_j^T$ in the other image, and is computed as:

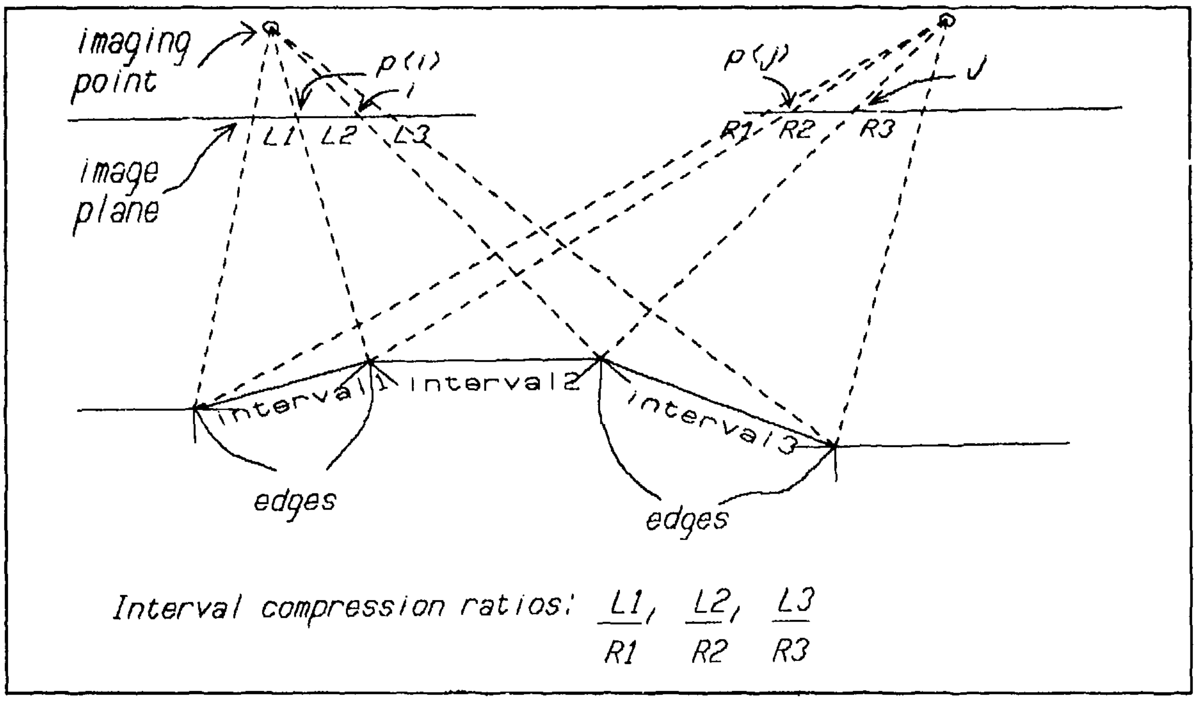
$$PRight^T_{ij} = GPROB(ER^T_{l,i} - ER^T_{r,j}), \qquad (4-8)$$

$$GPROB(x) = \int_{x-0.5}^{x+0.5} GPDF(0, \sigma^T_f), \qquad (4-9)$$

(*GPDF* being the Gaussian probability density function of (4-10) with parameters *mean* and *standard deviation*)

$$GPDF(\eta, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-0.5\left(\frac{(x-\eta)}{\sigma}\right)^2} \qquad (4-10)$$

$$PContrast^T_{ij} = \begin{cases} 1.0, & if\ Contrast(L^T_i) = Contrast(R^T_j), \\ 0.0, & otherwise. \end{cases}$$



*Interval compression ratio*
Figure 4-5

The first three terms of this probability product, composed in $PStat^T_{ij}$ form a *static* probabilitistic measure that may be precomputed for any particular epipolar pairing (they can be determined *a priori* from the edge properties). The last term, $PInterval^T$, interval correspondence probability, must be determined dynamically at each decision point in the Viterbi correlation (it is an *a posteriori* measure, depending upon the interval choices available). This interval correspondence probability, $PInterval^T$, estimates the probability that the intervals between two pairs of matched edges are the projections of the same surface. Currently this is computed in one of two ways ··· the first being a rough intuitive approximation, the other based on a result of Arnold. There seems to be very little difference in the results of the processing with these measures; Arnold's technique has only been introduced in this work quite recently, and the difference between the two has not as yet been fully explored.

In the rough approximation:

$$PInterval^T_{i,p(i),j,p(j)} = 0.75\left(1.0 - \left(\frac{\min(Length^T_{i,p(i)}, Length^T_{j,p(j)})}{\max(Length^T_{i,p(i)}, Length^T_{j,p(j)})}\right)^2\right)$$

where

$$Length^T_{m,n} = \begin{cases} Coordinate(R^T_m) - Coordinate(R^T_n), & \text{if Right image interval} \\ Coordinate(L^T_m) - Coordinate(L^T_n), & \text{if Left image interval} \end{cases}$$

From [Arnold 1980], the probability, based on an assumption of uniformly distributed surface orientations, has the cumulative distribution function $CDF$,

$$CDF = \int_{-\infty}^{\infty} \tan^{-1} \frac{1}{\frac{a}{R-1} + b}$$

where $R = \frac{Length^T_{i,p(i)}}{Length^T_{j,p(j)}}$,
$a = \frac{B}{z}$
$b = \frac{x}{z}$
$B =$camera baseline
$z =$scene distance,
and $x =$ lowest coordinate of edge in left image space.

Rather than integrating this probability density function, Arnold uses evaluations of the integrand over a uniformly distributed domain as his probabilistic measures.

It should be noted that the static probabilitistic measure $PStat^T_{ij}$ calculation would lead to a computation of $O(n^2)$, while the use of interval correspondence probability $PInterval$ brings the computation up to $O(n^3)$.

### *The Full Resolution Correspondence Process*

For full resolution matching, each edge is treated as a doublet, being a **left** half and a **right half.** The probability that one side of left image edge $L_i$ corresponds to the same side of right image edge $R_j$ is estimated as:

$$PFull_{ij} = PStat_{ij} \times PInterval_{i,p(i),j,p(j)} \qquad (4-11)$$

with $PStat_{ij} = PSid_{ij} \times POrient_{ij} \times PReducedRelDisp_{ij}$,

and $PInterval_{i,p(i),j,p(j)}$ defined as above.

$$PSid_{ij} = \begin{cases} \text{if } \textit{left} \text{ half of the edge, then the probability that the } \textit{left} \text{ intensity value of edge } L_i \\ \qquad \text{corresponds to the } \textit{left} \text{ intensity value of edge } R_j, \ (PLeft^0_{ij}, \text{ of (4-7)}), \\ \text{if } \textit{right} \text{ half of the edge, then the probability that the } \textit{right} \text{ intensity value of edge} \\ \qquad L_i \text{ corresponds to the } \textit{right} \text{ intensity value of edge } R_j, \ (PRight^0_{ij}, \text{ of (4-8)}). \end{cases}$$

$POrient_{ij}$ = probability that the orientation of edge $L_i$ corresponds to the orientation of edge $R_j$ in the other image.

This probability of edges corresponding based upon their image-plane orientation has been derived in two ways, as before. The first (ad hoc) was to determine a probabilistic weighting:

$$POrient_{ij} = 0.75(1-\epsilon^2), \qquad (4-12)$$

$$\epsilon = \frac{2min(Orientation(L_i), Orientation(R_j))}{\pi max(Orientation(L_i), Orientation(R_j))}.$$

where the factor 0.75 makes the probability integrate to 1.0. The other derivation comes from considering the probability of correspondence of two edges $L_i$ and $R_j$ as a bivariate distribution in $Orientation(L_i)$ and $Orientation(R_j)$ with the probability density function as depicted in Figure 4-6 (after Arnold).

$$PReducedRelDisp_{i,j} = 0.75(1-(Normdev^T)^2),$$

$Normdev^T$ = normalized deviation from reduced resolution interval disparity.

This latter component provides a *bias* from the disparities set by the reduced resolution correspondence process. It gives a bias toward edge pairings whose disparity is near that of their interval as a whole. Consider a potentially corresponding edge pair $(L_i, R_j)$, as depicted in Figure 4-7. The disparity associated with this pair matching is $Disp_{i,j} = Coordinate(L_i) - Coordinate(R_j)$. If the two edges come from a particular reduced resolution interval $Interval^T_{m,p(m),n,p(n)}$, whose average disparity is $ADisp^T_{m,p(m),n,p(n)}$, where:

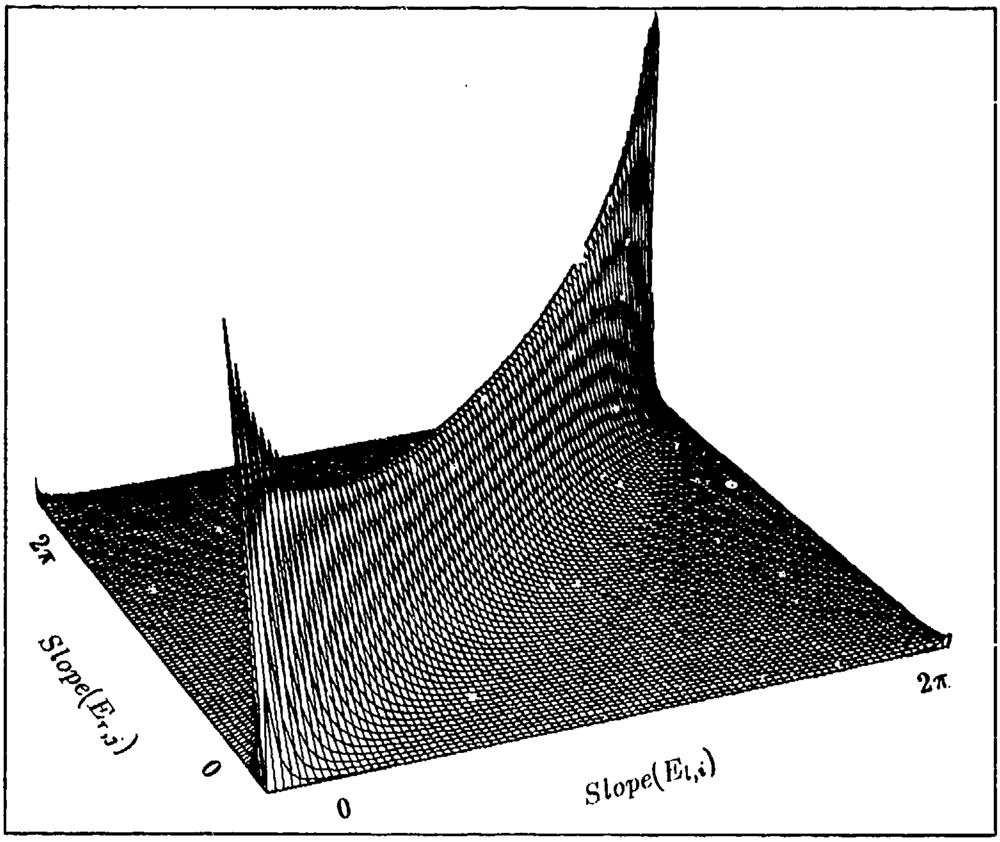
$$ADisp^T_{m,p(m),n,p(n)} = \frac{Disp^T_{m,n} - Disp^T_{p(m),p(n)}}{2}, \text{ and}$$

$$L^T_{p(m)} \le L_i \le L^T_m \text{ and } R^T_{p(n)} \le R_j \le R^T_n,$$

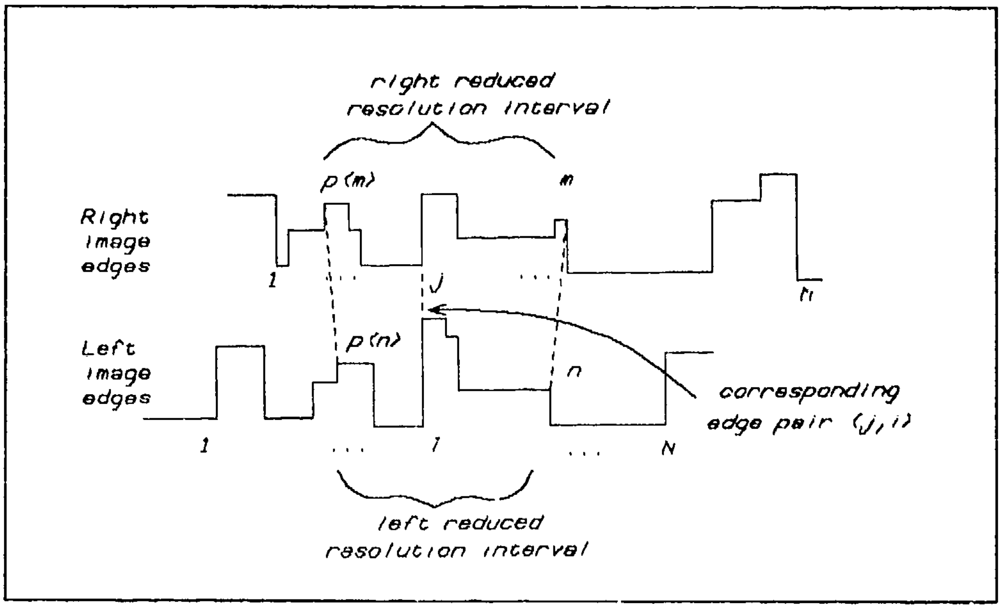
then the deviation in disparity

$$Ddev^T = ADisp^T_{m,p(m),n,p(n)} - Disp_{i,j} \qquad (4-13)$$

biases the probability of the edges $L_i$ and $R_j$ corresponding. The normalization is with respect to the size of the interval. Having not made an analysis of the distribution of this *bias* parameter, I use it as a probabilistic weighting $1-\epsilon^2$.



*Edge angle probability density function (from [Arnold 1980])*
Figure 4-6


right reduced resolution interval
p(m)
m
Right image edges
1
...
j
...
n
Left image edges
p(n)
n
corresponding edge pair (j, i)
1
...
i
...
N
left reduced resolution interval

*Reduced resolution disparity bias*
Figure 4-7

### *The Constrained-Interval Edge Correspondence Process*

The line-by-line constrained-interval edge matching, that which follows the cooperative continuity process, uses an evaluation function nearly identical to that used by the full resolution edge matching:

$$PInterE_{ij} = PStat_{ij} \times PInterval_{i,p(i),j,p(j)} \tag{4-14}$$

with $PStat_{ij} = PSid_{ij} \times POrient_{ij} \times PInterRelDisp_{ij}$,
and $PInterval_{i,p(i),j,p(j)}, PSid_{ij}$, *and* $POrient_{ij}$ defined as above.
$PInterRelDisp_{ij} = 0.75\Big(1 - (Normdev)^2\Big)$,
and $Normdev$ = normalized deviation from full resolution interval disparity.

## *4.1.3 – Intensity-based correspondences*

### *The Constrained-Interval Intensity Correspondence Process*

The line-by-line constrained-interval intensity matching, occuring only after the constrained-interval edge correspondence process, draws again on the measured pixel intensity variance. Here, the probability that pixel $Pixel_{l,i}$ in one image corresponds to pixel $Pixel_{r,j}$ in the other image is set as:

$$PPixel_{ij} = PIntensity_{ij} \times PLinearInterpolate_{ij} \tag{4-15}$$

where $PIntensity_{ij}$ = the integral of the Gaussian probability density function (zero-mean, $\sigma = \sigma_d$), about the intensity difference.

$$PIntensity_{ij} = \int_{\delta-0.5}^{\delta+0.5} GPDF(0, \sigma_d), \tag{4-16}$$

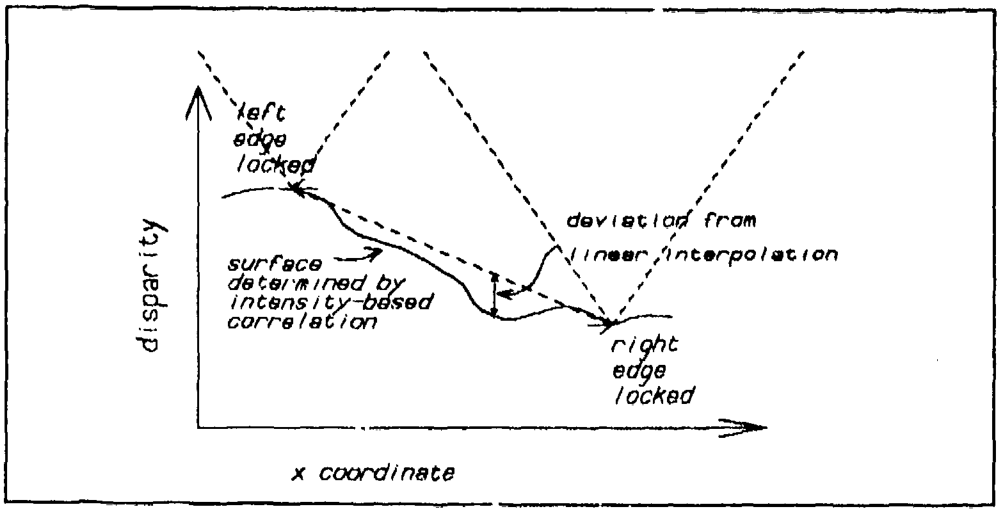
$$\delta = Intensity(Pixel_{l,i}) - Intensity(Pixel_{r,j}).$$

and

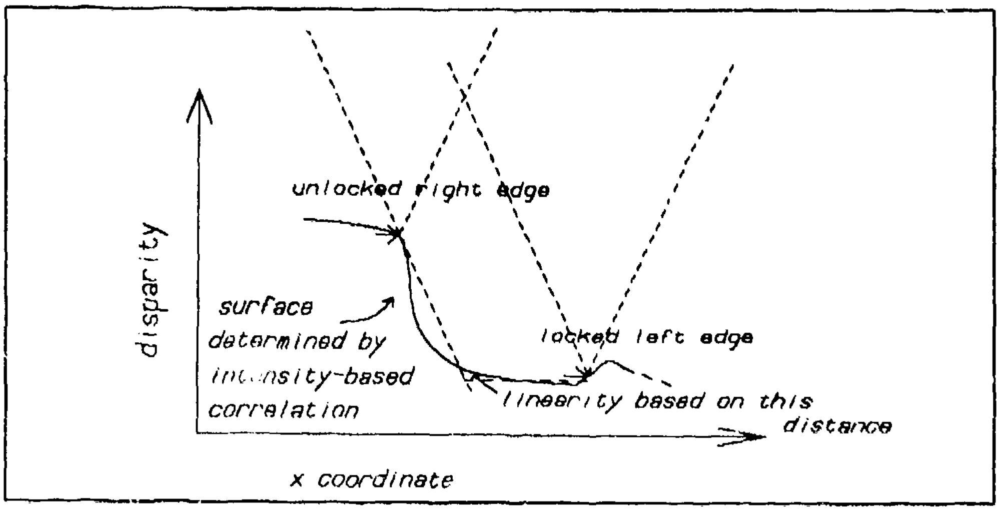
$$PLinearInterpolate_{ij} = 0.75(1 - \epsilon^2), \tag{4-17}$$

*$\epsilon$ = normalized deviation in disparity $Disp_{ij}$ from a linear interpolation over the interval in which the pixels occur.*

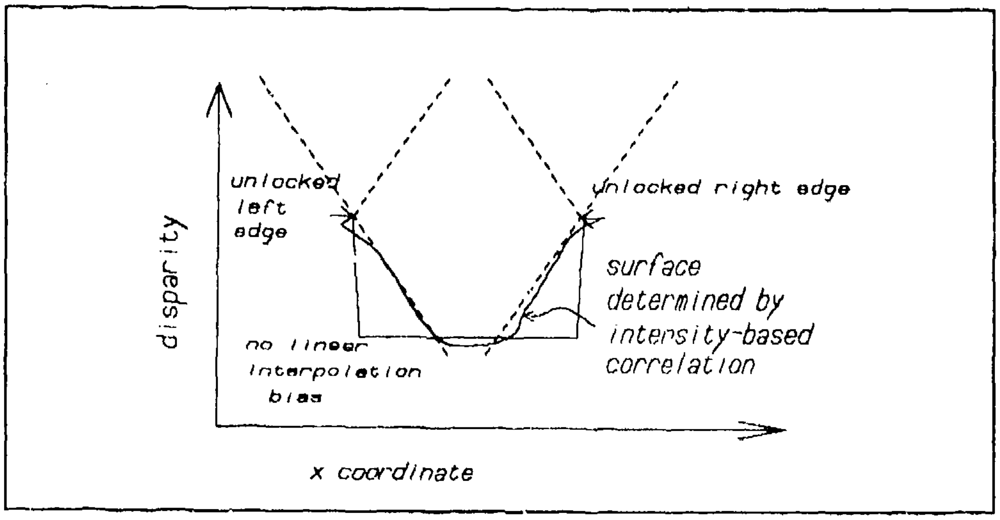
Recall that the edge-based matchings treat edges as doublets, being a left and a right half. Each of these halves has, independently, the possibility of matching a corresponding edge half in the other image. If it does match one, it is said to be *locked* at that point, otherwise it is *free*. Consider that an interval is locked on its left side by the right half of its leftmost edge, and on the right side by the left half of its rightmost edge. For the linear interpolation, if both sides of the interval are *locked*, then the deviation from a linear interpolation at a particular pixel pairing is the difference between its calculated disparity $Disp_{ij}$ and the associated interpolated disparity at that point. Figure 4-8 shows this situation. Figure 4-9 indicates the means for determining the deviation when one side of the interval is *free* (fails to be *locked*). If both are *free*, as shown in Figure 4-10, then $\epsilon = 0$.



**Intensity matching with both end points locked**
**Figure 4-8**



**Intensity matching with one end point free**
**Figure 4-9**



**Intensity matching with both end points free**
Figure 4-10

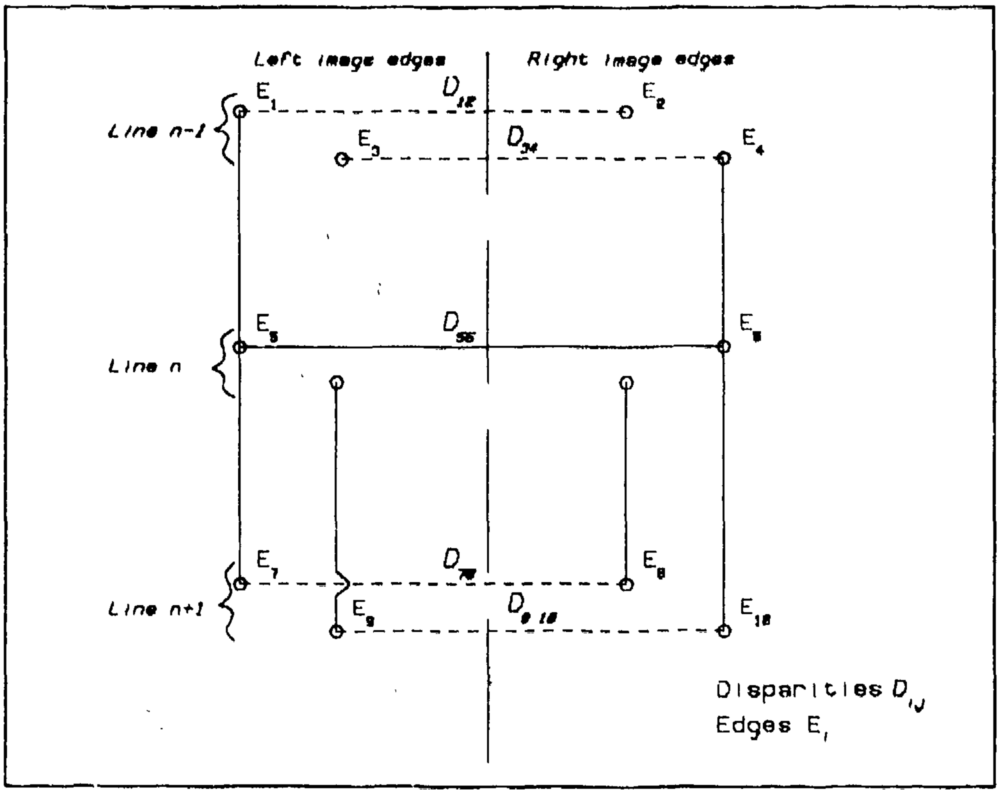
### 4.1.4 – Summary of correlation statistics

The two types of correlation statistics used in the processing are:

- intensity based, and using the Gaussian probability density function to estimate the likelihood of edges from opposite images matching, and
- geometrically based, ,where estimates of the distribution of scene characteristics are used to specify edge correspondence probabilities.

## 4.2 Cooperative Continuity Constraint Statistics

The cooperative continuity constraint process is also statistic driven. Each edge in an image has two-dimensional connectivity to the edges to which it is proximal (see Figure 4-11). While the full resolution correspondences are being formed, measures of the variation in disparity $Disp_{ij}$ between connected edges are made and accumulated to give a mean and standard deviation $[\mu, \sigma_{Disp}]$ of these inter-edge disparity differences. What these differences measure is the implied change in depth along the connected sequence of edges. Clearly these changes should be small along a continuous three-space curve. The accumulated disparity difference statistics $[\mu, \sigma_{Disp}]$ provide a metric for distinguishing between the good and the questionable correspondences chosen by the Modified Viterbi correlation — those disparity differences which lie outside of the $[\mu - \sigma_{Disp}, \mu + \sigma_{Disp}]$ difference window suggest abrupt changes in depth, discontinuities in the supposed continuous 3-space curves giving rise to the involved image edges. Reasons will be given in section 6-3 for a more arbitrary setting of $\sigma_{Disp} = 1.0$.



*Edge connectivity structure*

**Vertical lines joining edges are image plane connectivity; horizontal lines mark edge pairings assigned by the correspondence process.**

Figure 4-11

*Chapter 5*

# THE MODIFIED VITERBI CORRELATION ALGORITHM

## 5.1 The Correspondence Problem

When I first looked at the computation task of matching edges from one image with those in the other image, I thought in terms of having a heuristically bounded search which would optimize some metric. The combinatorics of matching $m$ edges from an epipolar line of one image with $n$ edges from the corresponding epipolar line in the other image, allowing for strictly one-to-one matching but not considering other distinguishing characteristics, is of order $(min(m,n))!$, which, for $m = n$ is $m!$. For a typical line of the Control Data Corporation imagery, $m = n = 11$, and $11! = 39,916,800$. A typical line of the Night Vision Laboratory imagery has $m = n = 30$, and $30! > 2.65 \times 10^{32}$. Obviously the combinatorics are rather overwhelming, and I put a lot of effort into analysis and design, trying to find methods to reduce or bound this cost. Certain obvious approaches come to mind, and these were implemented and tested:

- using a hierarchic scheme, where edge strength is used to order combinations, and correspondences are only allowed within strength intervals;
- precluding edges of differing contrasts from corresponding;
- limiting disparity values to a certain range;
- using a coarse to fine strategy, reducing image resolution to enable working first with the fewer reduced resolution edges.

In the interests of both parallelism and robustness, it was critical for the design that the results of the stereo matching be independent for each line processed (in contrast with the algorithm used in [Henderson 1979]), so I could not allow the solution from line $j$ to affect the order or results for the processing of scanline $j + 1$ (or $j - 1$ or $m \neq j$ for that matter), and this was one common heuristic that had to be avoided.

Accompanying these processing constraints was a quite involved evaluation function capable of estimating the maximum score attainable for the correlation from a particular set of correspondences. This use of an evaluation function estimator allowed the introduction of the extensive pruning of a branch and bound algorithm. Even with it, though, runs for certain lines took near minutes (on a DEC KL-10). A better approach was needed, and it appeared in a dynamic programming variant called the Viterbi algorithm.

The Viterbi algorithm is defined as a recursive optimal solution to the problem of estimating the state sequence of a discrete-time finite-state Markov process observed in memoryless noise ([Forney 1973]). The underlying Markov process is characterized as follows:

Time is discrete

The state $x_m$ at time $m$ is one of a finite number $N$ of states $n$, $1 \leq n \leq N$; *i.e.* the state space $X$ is simply $\{1, 2, \ldots, N\}$.

Assuming the process runs in time domain $T$ where $t \in [1, M]$, and the initial and final states $x_1$ and $x_M$ are known, the state sequence, mapping $T \rightarrow X$, can be represented as a vector $S = (x_1, x_2, \ldots, x_M)$.

The process is Markov in the sense that the probability $P(x_{m+1} \mid x_1, x_2, \ldots, x_m)$ of being in state $x_{m+1}$ at time $m + 1$, given all states up to time $m$ depends only on the state $x_m$ at time $m$:

$$P(x_{m+1} \mid x_1, x_2, \ldots, x_m) = P(x_{m+1} \mid x_m), \text{ and}$$

$$P(S) = \prod_{1 \leq i < M} P(x_{i+1} \mid x_i). \qquad (5-1)$$

In the problem addressed here of finding the optimal solution to the matching of edges from the left and right images, corresponding to the state space $X$ is the set of left image edges (numbered 1 to $N$ along a particular epipolar line); corresponding to the time domain $T$ is the set of right image edges (numbered 1 to $M$ along the corresponding epipolar line). The state sequence can be represented as a mapping:

$$f : T \rightarrow X,$$

or as a vector:

$$S = \{(m, n) \mid (m \overset{c}{\mapsto} n), m \in T, n \in X\},$$

$\overset{c}{\mapsto}$ is a binary relation indicating that $m$ in $T$ corresponds to $n$ in $X$

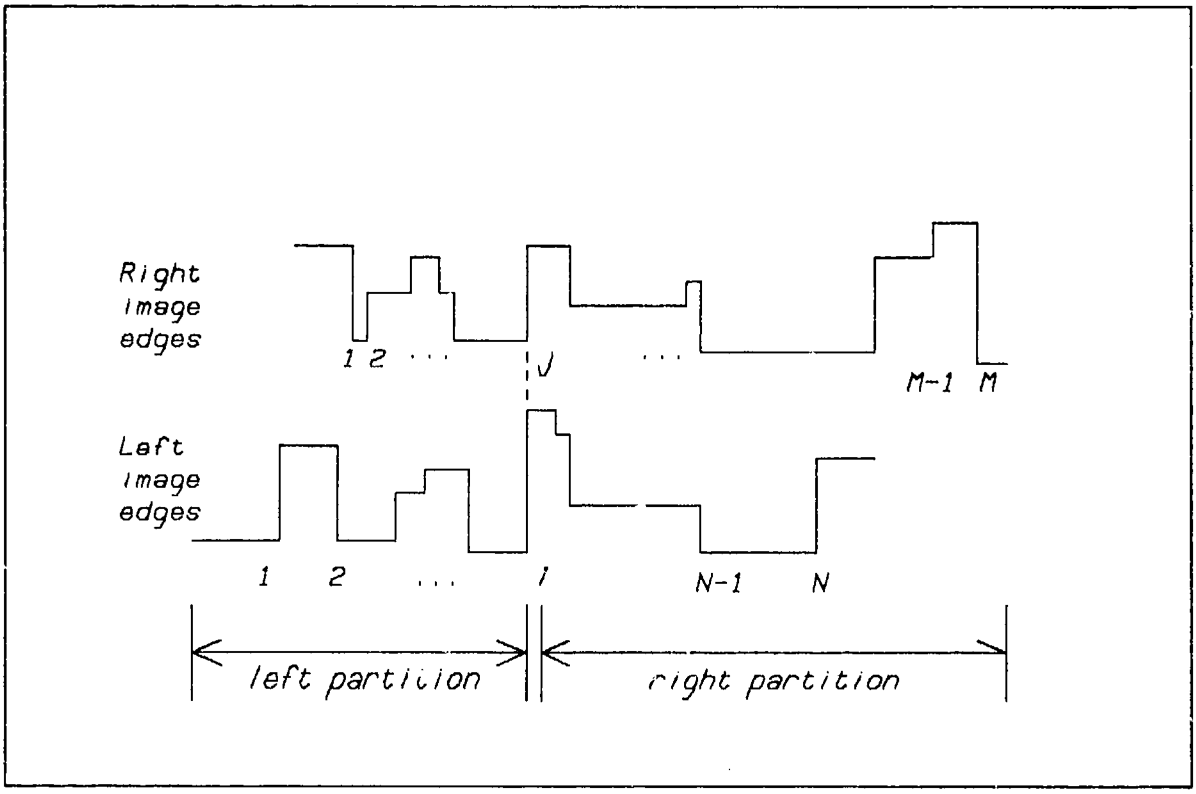
Regardless of representation, it is the record of correspondences for the various edges in $T$.

## 5.2 Direct Implementation of the Viterbi Algorithm

One of the assumptions capitalized on in the branch and bound scheme mentioned above held that there could be no edge reversals in the image plane. This meant that an edge sequence $L_i, L_j$ in one image, with $i < j$, and $i, j$ being edge indices, could not correspond to an edge sequence $R_k, R_l$ in the other image, if $k > l$ (refer to Figure 3-6). This is the *edge reversal* constraint, and was integral to the pruning. As it happens, this same constraint is the key to the use of the Viterbi algorithm.[17] It provides a **monotonicity** condition satisfying the sequencing constraint in the finite-state correspondence process. Consider Figure 5-1. What distinguishes the Viterbi technique from normal search is the ability to partition the original problem into two subproblems, each of which can be solved optimally and whose results can be processed to yield a global optimum for the original problem ('optimal' with respect to an evaluation function on the chosen parameters). In a recursive way, each of the subproblems may be divided and the solution process repeated. In particular, one can partition the problem of assigning correspondences among two tuples of edges $Edgeset_l$ and $Edgeset_r$ about some tentative pairing $(R_j \overset{c}{\mapsto} L_i)$, solve the associated correspondence problems of edges lying to the left of $L_i$ in $Edgeset_l$ with those lying to the left of $R_j$ in $Edgeset_r$ and edges

[17] [Rubin 1980] describes an image processing search technique, the Locus search, which is based on the Viterbi algorithm.

lying to the right of $L_i$ in *Edgeset*$_l$ with those lying to the right of $R_j$ in *Edgeset*$_r$. ( $\overset{c}{\mapsto}$ represents a binary relation, and ($a \overset{c}{\mapsto} b$) is read ***"a in Edgeset$_r$ corresponds to b in Edgeset$_l$".***)



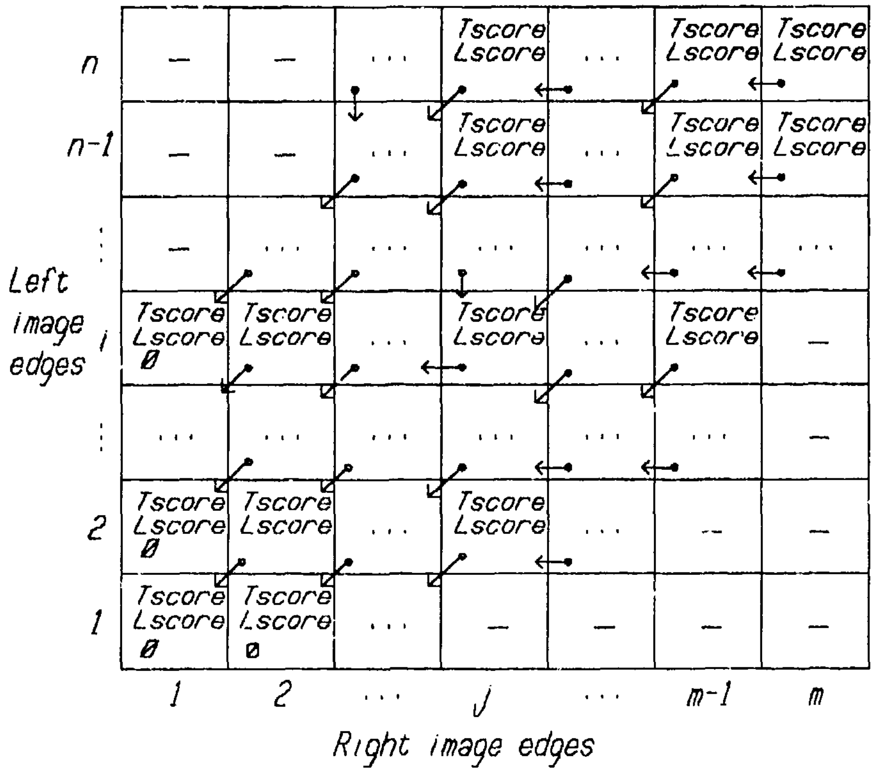
*Typical Right and Left Image Corresponding Epipolar Line Edges*

$$Edgeset_l = (L_1, L_2, \ldots, L_i, \ldots, L_{N-1}, L_N),$$

$$Edgeset_r = (R_1, R_2, \ldots, R_j, \ldots, R_{M-1}, R_M)$$

Figure 5-1

The optimal solution for the line correlation is that sequence of edge pairings from the left and right image lines which is consistent with this monotonicity constraint and maximizes some score. The score used here was based on summing the individual probability measures for each possible edge-pair correspondence (4-11). This summing favours the densest possible surface intrepretation ([Julesz 1976]). Other scorings, such as normalizing, summing weighted probability contributions, or taking the (more standard) product of probabilities (as defined in (5-3)) do not support this density preference. Consider a two dimensional array *Parray*$[1 : N, 1 : M]$ with *Edgeset*$_r$ along the bottom axis, and *Edgeset*$_l$ up the left side axis, as in Figure 5-2. The Viterbi solution implemented here develops from the left of the image (***right*** image edge index of 1) to the right of the image (***right*** image edge index of $M$), and within this, from the bottom (*left* image index of 1) to the top (***left*** image index of $N$). The first set of subproblems is all those involving the assignment of $R_1$. The second set of subproblems deals with $R_2$ using the results of the analysis of $R_1$. Thus for $M$ edges in the right image line, there are $M$ subproblem sets. A useful mnemonic to bear in mind about this processing is to ask, at each possible pairing ($R_j \overset{c}{\mapsto} L_i$), "***what is the best possible solution to the left of*** $(i, j)$ *if* ($R_j \overset{c}{\mapsto} L_i$)". The set of solutions (including the optimal) is built up by evaluating this for all $(i, j)$.



*Viterbi Dynamic Programming Array*
Figure 5-2

The matching process is monotonic in both left image edge indices $\{i \mid i \in [1,N], L_i \in \boldsymbol{Edgeset}_l\}$, and right image edge indices $\{j \mid j \in [1,M], R_j \in Edgeset_r\}$. This monotonicity means that the solution for the pairing $(R_j \overset{c}{\mapsto} L_i)$ need only examine that portion of $\boldsymbol{Parray}[1:N, 1:M]$ where $n \le i, m \le j$, *i.e.* the rectangular subarray whose top right corner is $\boldsymbol{Parray}[i,j]$ (otherwise, say if $(R_{j+1} \overset{c}{\mapsto} L_{i-1})$ preceded $(R_j \overset{c}{\mapsto} L_i)$, we note that $j+1 > j$ and the monotonicity is violated). The solution for the pairing $(i,j)$ is the best assignment of edges from $Edgeset_{l,p}, p \in [1,i]$ and $Edgeset_{r,q}, q \in [1,j]$, that is, for the edges in the two sets up to and including edges $L_i$ and $R_j$. A scoring function is defined for the various transitions possible in the processing, and these can usually be limited (because of the monotonicity) to the obvious three:

$$\{(\delta_x, \delta_y) = (-1,0), (-1,-1), (0,-1)\}. \tag{5-2}$$

Through this, subproblem $(Edgeset_{l,i}, Edgeset_{r,j})$ can be solved after subproblem $(Edgeset_{l,i-1}$, $Edgeset_{r,j})$ and subproblem $(Edgeset_{l,i}, Edgeset_{r,j-1})$ are solved (these both imply the solution of subproblem $(Edgeset_{l,i-1}$, $Edgeset_{r,j-1}))$. Thus, the decision for any pairing $(i,j)$ can be made with just 3 scoring comparisons, making the total line correlation computation $O(MN)$ (or $O(N^2)$, where $M$ and $N$ are of the same order).

An entry in $Parray[i,j]$ has associated with it:

- a local score, *Localscore*,
- a cumulative score (from the left), *Totalsccre*, and
- an indicator of the pairing $(k,l), k \leq i, l \leq j$, that is the predecessor to $(i,j)$ in the solution that contains $(R_j \overset{c}{\mapsto} L_i)$.

Each such entry in *Parray* is linked to other entries in *Parray* via these predecessor indicators. A chain of these entries contains a locally optimal solution to the line correspondence problem. The optimal chain over this entire set of chains is the global optimum for the whole line correlation (**note:** the chain of the best solution will begin with an entry in column $M$, specifically, the highest scoring entry in that column).

## 5.3 Modifications to the Viterbi Algorithm

The preceding overview of the scoring mechanism has been slightly misleading, as it doesn't take into account several issues ... those which relate to specific aspects of the various correspondence processes to be performed. The four correspondence processes — *reduced resolution edge, full resolution edge, constrained-interval edge*, and *constrained-interval intensity* — each have characteristics which make the above general outline inappropriate. The principal variation comes from:

- the treatment of unassignable pixels or edges (those which may be *obscured* in the other image, or be merely *spurious*).

  The complication this introduces is apparent when looking at the optimization metric used by the Viterbi method. Probabilities are treated multiplicatively (5-3). If one of the right image edges, $R_j$ has no correlate in the left image, then the optimal solution should have $P(R_j \overset{c}{\mapsto} L_i) = 0, \forall i \in [1,N]$. But even a single zero probability will take the total probability product to zero, since

$$P(S) = \prod_{1 \leq m < M} P(x_{m+1} \mid x_m) = 0, \text{ if } P(x_i \mid x_{i-1})) = 0, \text{ for } i \leq M. \qquad (5-3)$$

  Viterbi was not designed with *time* domain skipping in mind (although having a particular *state* unused would present no problem). The scoring mechanism must allow unmatched pixels and edges in both the left and the right images.

Two other issues also affect the implementation of the Viterbi algorithm. These are:

- the metrics used in the scoring. One, *interval compression ratio*, drives the computation to $O(N^3)$.

  The information needed to compute the left and right image interval ratio will (in general) not be available to the local $(\delta_x, \delta_y)$ transition rule of (5-2). It is conceivable, especially when considering the possibilities of the prior unassigned pixel or edge variation, that this computation may need to look as far back as the first left image edge! The transition mechanism should minimize this search while maintaining optimality.
- the edge index numbering conventions (in certain cases each edge is considered a doublet — its left and right sides).

With edges split into a left and a right half, the computation is increased (proportional to the order), and the $(\delta_x, \delta_y)$ transition mechanism (5-2) may need to be altered. The efficiencies possible in that the left sides of edges cannot match the right sides of edges should be used in reducing the increased combinatorics arising from the edge splitting.

Considering these difficulties one at a time, the variations they introduce are as follows:

[*Unassignable edges*] Edges from either image that are either ***spurious***, or are ***obscured*** in the other image, should be left unassigned by the matching process. This means that chains of pairings (in the various solutions) may not be joining adjacent edges — there must be provision for skipping over certain (unassigned) edges in these chains. This is accomplished by allowing an edge $R_j$ to match the null edge $L_{i\phi}$. The alternative, not providing for the interpretation of certain edges as being spurious or obscured, is both unrealistic and unacceptable — there will always be edges which have either no visible correlate in the other image or no physical justification in the scene.

Unmatched pixels from the ***constrained-interval intensity*** correspondence process don't require such special-case treatment, as they are positioned by interpolation.

[*Scoring metrics*] ***Interval compression ratio*** is a measure of the perspective foreshortening of scene surfaces. This is recognized in the psychological literature as a cue to ***stereopsis*** ([Blakemore 1970]). Its computation here requires looking back from a pairing $(R_j \overset{c}{\mapsto} L_i)$ in ***Parray*** to the preceding edge pairing, and since this need not necessarily be an adjacent edge ($L_{i-1}$ or $R_{j-1}$), the entire incident subarray may need to be searched. In fact the algorithm can be structured such that the preceding column ($\boldsymbol{Parray}[n, j-1], n \in [1, i-1]$) is all that is required here. Nevertheless, this takes the computation for the three edge-based matchings to $O(N^3)$ from $O(N^2)$.

A very important implementation detail should be noticed here: to guarantee optimality in those cases where unassigned edges appear in the intervals considered **does**, in general, require an $N^2$ search over the preceding subarray, making the computation $O(N^4)$ where these occur. The problem is that when using interval compression ratio unassigned pairings of edge $R_j$ cannot make an optimal choice for their predecessor since, as indicated, the choice will depend upon the ***assigned*** pairing of some edge to the right of $R_j$. Savings can be made on this by maintaining lists of possible predecessors for each unassigned pairing $(R_j \overset{c}{\mapsto} L_{i\phi})$. Since only predecessor paths to assigned pairings will affect the decision, somewhat less search will be necessary in finding the optimal path (in the degenerate case this would still be $N^2$). A near-optimal solution is found here, where each unassigned pairing is forced to make a decision about its predecessor.

***Constrained-interval intensity*** matching is not edge-based, so uses quite different optimization metrics from those so far mentioned.

[*Edge numbering*] Each image edge in the ***full resolution*** matching is treated as a doublet, its left and its right sides. A left side of an edge can only match a left side of another edge, and a right side of an edge can only match a right side. This splitting allows contrast reversals to be handled correctly, occuring, for example, when a grey object is seen above a checker-board with the left image seeing it in relief against the white, and the right image seeing it in relief against the black. Psychological evidence suggests that human vision cannot achieve stereopsis under conditions of such contrast reversal whereas the algorithmic mechanisms in a computational vision scheme will enable edge matching here (this shows a situation where deviating from the characteristics of the human vision system allows a greater flexibility in the processing). Providing for this special edge treatment doubles the number of edges with which the system must deal, so multiplies the standard correlation computation of $O(N^3)$ by $2^2 = 8$ in time and $2^2 = 4$ in space. With the consideration of half-edge polarity, the increased time complexity is reduced from 8 to $2^2 = 4$.

The ***reduced resolution*** correspondence process doesn't allow contrast reversals, so doesn't have this accompanying increase in computation cost.

### *5.3.1 – Edge-based matching*

The chapter on statistics describes in fair detail the optimization metrics used in the edge-based matchings — it should be referred to for computational specifics of the general outline that follows here.

The ***reduced resolution*** correspondence process evaluates the matchings of reduced resolution edges on the basis of:

1) contrast about the edge,
2) intensity difference about the edge (both sides),
3) ***interval compression ratio*** between matched edges.

Since it does not allow contrast reversals, it does not treat edges as doublets. Rather, each edge enters the correlation only once, and the ***Parray*** has vertical indices $\{ i \mid i \in [1, N], L_i^T \in Edgeset_l^T \}$ and horizontal indices $\{ j \mid j \in [1, M], R_j^T \in \text{ʚ}dgeset_r^T \}$. The suggestion here is that it is the high-frequency components of the images which will exhibit this contrast variation, and the low-frequency components will be expected to be less varying. The intensity variation metric is the product of the probabilities that the *left* sides of the edges correspond and the *right* sides of the edges correspond, and this is just the product of the two integrals of the Gaussian probability density functions, as detailed in (4-9). ***Interval compression ratio*** usage means that this computation is $O(N^3)$ in edges. Since the ***reduced resolution*** correspondence process is so similar to the ***full resolution*** correspondence process, yet simpler in its handling of just single edges (as opposed to doublets), the example of the Viterbi algorithm correlation to be presented at the end of this chapter will detail only the processing of a full resolution line pair — the functioning of the reduced resolution correspondence process should be fairly obvious once the full resolution process is understood. Section 5-4 will present this processing example.

As mentioned, in the ***full resolution*** correspondence process, edges are treated as doublets, their **left** and **right** sides, so the matcher evaluates the correspondences of image *half-edges* on the basis of:

1) intensity difference at the appropriate side of the edge,
2) orientation of the edges,
3) ***interval compression ratio*** between this edge and its predecessor,
4) disparity ***bias*** as set by the ***reduced resolution*** correspondence process.

Each edge enters the correlation twice, giving the $\boldsymbol{Parray}$ vertical indices $\{\, i \mid i \in [2, 2N+1], L_{\lfloor \frac{i}{2} \rfloor} \in Edgeset_l \,\}$, and horizontal indices $\{\, j \mid j \in [2, 2M+1], R_{\lfloor \frac{j}{2} \rfloor} \in Edgeset_r \,\}$. The intensity variation metric measures the probability that the sides of the edges correspond, and this is just the integral of the Gaussian probability density function as detailed in (4-9). The orientations of the edges, as measured in roughly $n$ degree increments (determined by a scatter analysis) affect the optimization scoring as indicated in (4-12). (4-13) outlines the computation of the disparity ***bias*** measure. This, in conjunction with the search interval definition of the ***reduced resolution*** correspondence, constrains the range of choices in ***full resolution*** edge correspondences.[18] Again, the use of ***interval compression ratio*** means that the correlation computation is $O(N^3)$ in half-edges. The solution to the correspondence problem for edges along conjugate epipolar lines is specified by the set:

$$\{(j, i) \mid R_j \leftrightarrow L_i\}, \qquad (5-4)$$

where $L_i$ and $R_j$ refer to ***half-edges*** of the full-resolution correspondence process.

For ***constrained-interval*** edge matching, the results of the previous two (***reduced*** and ***full resolution***) correspondence processes have acted to associate together intervals along conjugate epipolar lines. The edges in these intervals which failed to find matches in the full correspondence process are re-examined to see whether the more tightly constraining context will now permit them to be matched across images. Edges are again treated as doublets, their **left** and **right** sides. Here, the correlator evaluates the correspondences of these image ***half-edges*** on the basis of:

1) intensity difference at the appropriate side of the edge,
2) orientation of the edges,
3) ***interval compression ratio*** between this edge and its predecessor,
4) disparity ***bias*** as set by the ***full resolution*** correspondence process.

Notice that these are almost identical to the optimization metrics used for ***full resolution*** matching. The difference is that in ***constrained-interval*** matching the ***bias*** measure is about the centre of the interval in which the edges find themselves after the ***full resolution*** correspondence process, rather than about the centre of the intervals defined by the ***reduced resolution*** matching.

### *5.3.2 – Intensity-based matching*

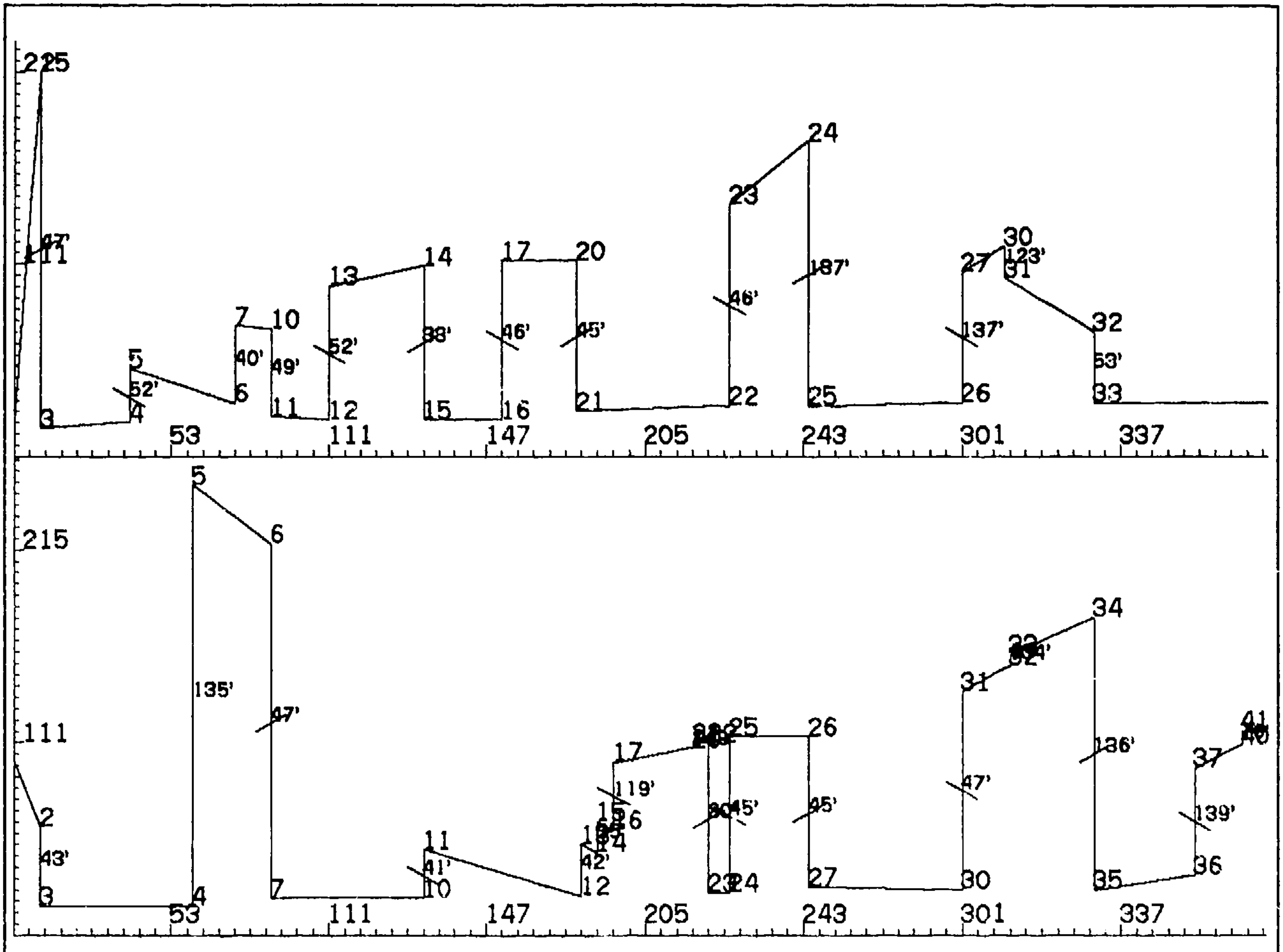
The ***constrained-interval*** intensity correspondence process finds the optimal correspondence of individual pixels. It looks at pixels in the intervals associated together by the ***reduced resolution***, ***full resolution*** and ***constrained-interval*** edge correspondence processes. (4-16) indicates the probabilistic measures used for this optimization. The matching of intensity values is a standard correlation technique, and its analysis is based on the image intensity variance statistics measured at the start of the processing. The role of the *PLinearInterpolate* metric (4-17) is a little less obvious. It functions to pull the implied surface toward a linear interpolation with end point conditions as indicated in Figures 4-8 through 4-10. Perhaps a better metric would be one which used a ***smoothing*** measure, looking for continuity in a few derivatives of the implied surface slope. Further refinements to this stereo process will include incorporation of an improved interpolation metric.

[18] It is interesting to note, as pointed out in [Schumer 1979], that low spatial frequency gratings can be fused at much larger disparities than can higher spatial frequency gratings.

## *5.4 A Line-Pair Viterbi Edge Correlation*

This section demonstrates the processing of the Viterbi algorithm on a pair of corresponding full resolution image lines. The line-pair are numbers $244_8$ of the Control Data images, and are seen in Figure 5-3 (edge indices are in octal). The possible right to left edge pairings, indicated by ***right*** : $left_1, left_2, \ldots, left_n;$, are as follows:

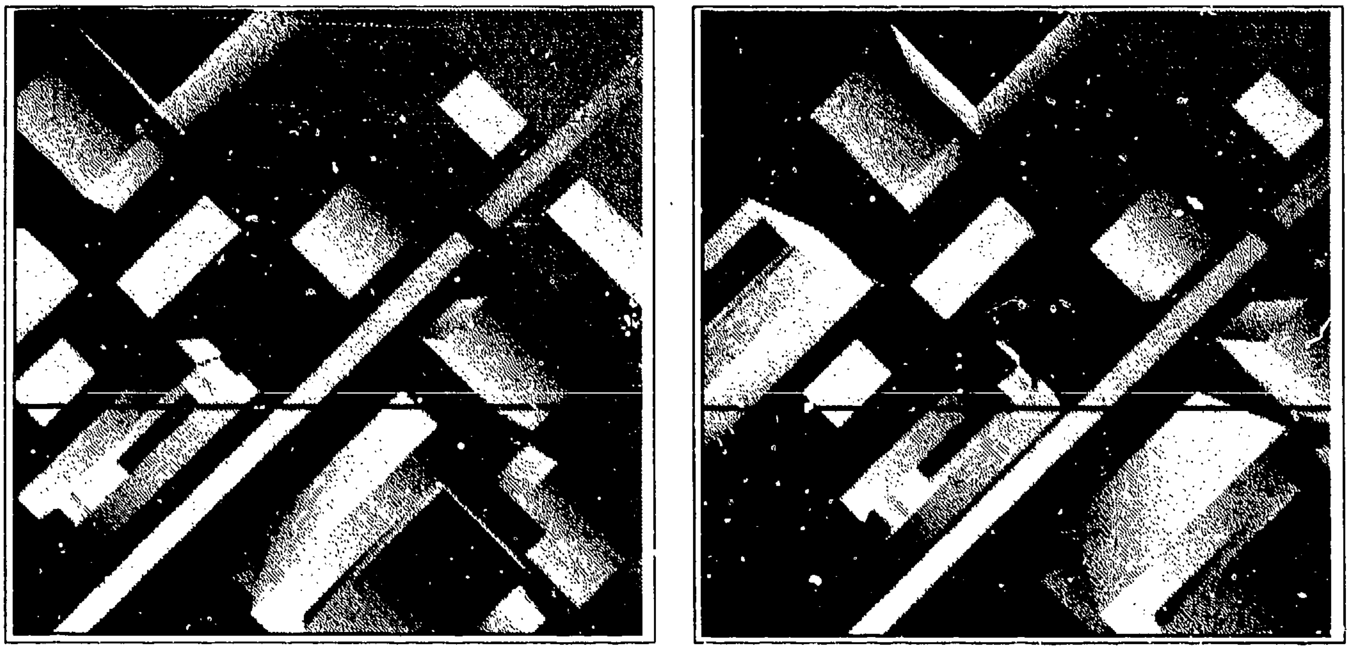
(**2** : 6; **3** : 3, 7, 11, 13, 15; **4** : 10, 12, 14; **5** : 7, 11, 13, 15; **6** : 10, 12, 14; **7** : 11, 13, 15; **10** : 12, 14, 22;
**11** : 7, 11, 13, 15; **12** : 10, 12, 14; **13** : 11, 13, 15; **14** : 22; **15** : 23; **16** : 24; **17** : 25; **20** : 26;
**21** : 27; **22** : 30; **23** : 31; **24** : 34; **25** : 35; **26** : 36; **27** : 37, 41; **30** : 40; **31** : 37, 41;)



*Edges of left and right image line 244*
Figure 5-3

The right image line is shown above its corresponding left image line. Half-edge indices increase from left to right — from 2 to 33 for the right image line and from 2 to 41 for the left image line (left half-edges are numbered even, and right half-edges are numbered odd). Figure 5-4 locates this line pair in the two images. The *reduced resolution* correlation for this pair of lines resulted in the edge correspondences indicated by the diagonal strokes between edge numbers (for example

(2/3, 6/7) correspond, as do (4/5, 10/11)). These matchings constrain the possible pairings of the ***full resolution*** correspondence process. (4-13) indicates a biasing mechanism affecting the probabilistic estimates for matching full resolution edges from corresponding intervals. The interval shown there would seem to be that delimited by nearest-neighbour matched edges (nearest-neighbour diagonal markings in the profiles of Figure 5-3) — in fact this interval constraint is loosened somewhat, and an interval is defined as the union of this interval and those neighbouring it. The biasing uses this broader range for its probability estimates, and only edges in such corresponding ***broader*** intervals are considered as candidates for matching in the full resolution correspondence process. The reason for this redefinition of intervals is that the reduced resolution correspondence process ***can*** make mistakes, and a little flexible interpretation is called for in using its suggested constraints). It could also be argued that a low to high resolution matching is not an adequate model for correspondence control, and again the broader scope diminishes the negative aspects of this strategy.[19]



*The horizontal mark in the images indicates line* $244_8$
Figure 5-4

A linked list depiction of the Viterbi array ***Parray***, Figure 5-5 below, contains all of the ***Localscore*** and ***Totalscore*** measures and the associated ***Predecessor*** for each possible ***full resolution*** edge pairing. The designation $-n$ for a left edge index indicates that the right edge is being considered as paired with the null edge $L_{n\phi}$. This should be interpreted as meaning that the right edge is spurious and temporarily positioned between $L_n$ and $L_{n+1}$, or is obscured from view from the left imaging point and again positioned between $L_n$ and $L_{n+1}$. Figure 5-6 shows the two dimensional structure of the ***Parray***, with the arrows indicating the predecessor links specified in Figure 5-5. The solution is marked in bold.

[19] The recent [Mayhew 1981] paper discusses more comprehensive control strategies.

| i:(R,L | *Total*, *Local* | Pre) |
|---|---|---|
| 1:(2,-2 | 0.000 | -) |
| 2:(2,-3 | 0.000 | -) |
| 3:(2,-4 | 0.000 | -) |
| 4:(2,-5 | 0.000 | -) |
| **5:(2,6** | 0.914, 0.914 | -) |
| 6:(2,-6 | 0.000 | -) |
| 7:(3,-2 | 0.000 | 1) |
| 8:(3,3 | 0.059, 0.059 | 1) |
| 9:(3,-3 | 0.000 | 2) |
| 10:(3,-4 | 0.000 | 3) |
| 11:(3,-5 | 0.000 | 4) |
| 12:(3,-6 | 0.914 | 5) |
| **13:(3,7** | 1.686, 0.771 | 5) |
| 14:(3,-7 | 0.914 | 5) |
| 15:(3,-10 | 0.914 | 5) |
| 16:(3,11 | 1.054, 0.140 | 5) |
| 17:(3,-11 | 0.914 | 5) |
| 18:(3,-12 | 0.914 | 5) |
| 19:(3,13 | 0.933, 0.019 | 5) |
| 20:(3,-13 | 0.914 | 5) |
| 21:(3,-14 | 0.914 | 5) |
| 22:(3,15 | 0.917, 0.003 | 5) |
| 23:(3,-15 | 0.914 | 5) |
| 24:(4,-6 | 0.914 | 12) |
| 25:(4,-7 | 1.686 | 13) |
| **26:(4,10** | 2.078, 0.393 | 13) |
| 27:(4,-10 | 1.686 | 13) |
| 28:(4,-11 | 1.686 | 13) |
| 29:(4,12 | 1.733, 0.048 | 13) |
| 30:(4,-12 | 1.686 | 13) |
| 31:(4,-13 | 1.686 | 13) |
| 32:(4,14 | 1.699, 0.013 | 13) |
| 33:(4,-14 | 1.686 | 13) |
| 34:(4,-15 | 1.686 | 13) |
| 35:(5,-6 | 0.914 | 24) |
| 36:(5,7 | 1.052, 0.138 | 24) |
| 37:(5,-7 | 1.686 | 25) |
| 38:(5,-10 | 2.078 | 26) |
| **39:(5,11** | 2.832, 0.754 | 26) |
| 40:(5,-11 | 2.078 | 26) |
| 41:(5,-12 | 2.078 | 26) |
| 42:(5,13 | 2.296, 0.217 | 26) |
| 43:(5,-13 | 2.078 | 26) |
| 44:(5,-14 | 2.078 | 26) |
| 45:(5,15 | 2.185, 0.107 | 26) |
| 46:(5,-15 | 2.078 | 26) |
| 47:(6,-6 | 0.914 | 35) |
| 48:(6,-7 | 1.686 | 37) |
| 49:(6,10 | 2.075, 0.389 | 37) |
| 50:(6,-10 | 2.078 | 38) |
| 51:(6,-11 | 2.832 | 39) |
| **52:(6,12** | 3.231, 0.398 | 39) |
| 53:(6,-12 | 2.832 | 39) |
| 54:(6,-13 | 2.832 | 39) |
| 55:(6,14 | 3.083, 0.250 | 39) |
| 56:(6,-14 | 2.832 | 39) |
| 57:(6,-15 | 2.832 | 39) |
| 58:(7,-6 | 0.914 | 47) |
| 59:(7,-7 | 1.686 | 48) |
| 60:(7,-10 | 2.078 | 50) |
| 61:(7,11 | 2.372, 0.297 | 49) |
| 62:(7,-11 | 2.832 | 51) |
| 63:(7,-12 | 3.231 | 52) |
| **64:(7,13** | 3.636, 0.405 | 52) |
| 65:(7,-13 | 3.231 | 52) |
| 66:(7,-14 | 3.231 | 52) |
| 67:(7,15 | 3.721, 0.490 | 52) |
| 68:(7,-15 | 3.231 | 52) |
| 69:(10,-6 | 0.914 | 58) |
| 70:(10,-7 | 1.686 | 59) |
| 71:(10,-10 | 2.078 | 60) |
| 72:(10,-11 | 2.832 | 62) |
| 73:(10,12 | 2.854, 0.022 | 62) |
| 74:(10,-12 | 3.231 | 63) |
| 75:(10,-13 | 3.636 | 64) |
| **76:(10,14** | 3.770, 0.133 | 64) |
| 77:(10,-14 | 3.636 | 64) |
| 78:(10,-15 | 3.721 | 67) |
| 79:(10,-16 | 3.721 | 67) |
| 80:(10,-17 | 3.721 | 67) |
| 81:(10,-20 | 3.721 | 67) |
| 82:(10,-21 | 3.721 | 67) |
| 83:(10,22 | 3.736, 0.014 | 67) |
| 84:(10,-22 | 3.721 | 67) |
| 85:(11,-6 | 0.914 | 69) |
| 86:(11,7 | 0.981, 0.067 | 69) |
| 87:(11,-7 | 1.686 | 70) |
| 88:(11,-10 | 2.078 | 71) |
| 89:(11,11 | 2.244, 0.165 | 71) |
| 90:(11,-11 | 2.832 | 72) |
| 91:(11,-12 | 3.231 | 74) |
| 92:(11,13 | 3.563, 0.332 | 74) |
| 93:(11,-13 | 3.636 | 75) |
| **94:(11,-14** | 3.770 | 76) |
| 95:(11,15 | 3.886, 0.117 | 76) |
| 96:(11,-15 | 3.770 | 76) |
| 97:(11,-16 | 3.770 | 76) |
| 98:(11,-17 | 3.770 | 76) |
| 99:(11,-20 | 3.770 | 76) |
| 100:(11,-21 | 3.770 | 76) |
| 101:(11,-22 | 3.770 | 76) |
| 102:(12,-7 | 1.686 | 87) |
| 103:(12,10 | 1.772, 0.086 | 87) |
| 104:(12,-10 | 2.078 | 88) |
| 105:(12,-11 | 2.832 | 90) |
| 106:(12,12 | 3.377, 0.545 | 90) |
| 107:(12,-12 | 3.231 | 91) |
| 108:(12,-13 | 3.636 | 93) |
| 109:(12,14 | 3.689, 0.053 | 93) |
| **110:(12,-14** | 3.770 | 94) |
| 111:(12,-15 | 3.886 | 95) |
| 112:(12,-16 | 3.886 | 95) |
| 113:(12,-17 | 3.886 | 95) |
| 114:(12,-20 | 3.886 | 95) |
| 115:(12,-21 | 3.886 | 95) |
| 116:(12,-22 | 3.886 | 95) |
| 117:(13,-10 | 2.078 | 104) |
| 118:(13,11 | 2.082, 0.004 | 104) |
| 119:(13,-11 | 2.832 | 105) |
| 120:(13,-12 | 3.377 | 106) |
| 121:(13,13 | 3.467, 0.090 | 106) |
| 122:(13,-13 | 3.636 | 108) |
| 123:(13,-14 | 3.770 | 110) |
| **124:(13,15** | 4.006, 0.236 | 110) |
| 125:(13,-15 | 3.886 | 111) |
| 126:(13,-16 | 3.886 | 112) |
| 127:(13,-17 | 3.886 | 113) |
| 128:(13,-20 | 3.886 | 114) |
| 129:(13,-21 | 3.886 | 115) |
| 130:(13,-22 | 3.886 | 116) |
| 131:(14,-21 | 4.006 | 124) |
| **132:(14,22** | 4.148, 0.143 | 124) |
| 133:(14,-22 | 4.006 | 124) |
| 134:(15,-22 | 4.148 | 132) |
| **135:(15,23** | 4.268, 0.120 | 132) |
| 136:(15,-23 | 4.148 | 132) |
| 137:(16,-23 | 4.268 | 135) |
| **138:(16,24** | 4.474, 0.205 | 135) |
| 139:(16,-24 | 4.268 | 135) |
| 140:(17,-24 | 4.474 | 138) |
| **141:(17,25** | 5.417, 0.943 | 138) |
| 142:(17,-25 | 4.474 | 138) |
| 143:(20,-25 | 5.417 | 141) |
| **144:(20,26** | 6.359, 0.941 | 141) |
| 145:(20,-26 | 5.417 | 141) |
| 146:(21,-26 | 6.359 | 144) |
| **147:(21,27** | 7.326, 0.967 | 144) |
| 148:(21,-27 | 6.359 | 144) |
| 149:(22,-27 | 7.326 | 147) |
| **150:(22,30** | 8.249, 0.922 | 147) |
| 151:(22,-30 | 7.326 | 147) |
| 152:(23,-30 | 8.249 | 150) |
| **153:(23,31** | 9.032, 0.783 | 150) |
| 154:(23,-31 | 8.249 | 150) |
| 155:(24,-31 | 9.032 | 153) |
| 156:(24,32 | 9.142, 0.110 | 153) |
| 157:(24,-32 | 9.032 | 153) |
| 158:(24,-33 | 9.032 | 153) |
| **159:(24,34** | 9.498, 0.466 | 153) |
| 160:(24,-34 | 9.032 | 153) |
| 161:(25,-34 | 9.498 | 159) |
| **162:(25,35** | 10.279, 0.781 | 159) |
| 163:(25,-35 | 9.498 | 159) |
| 164:(26,-35 | 10.279 | 162) |
| **165:(26,36** | 10.768, 0.489 | 162) |
| 166:(26,-36 | 10.279 | 162) |
| 167:(27,-36 | 10.768 | 165) |
| **168:(27,37** | 11.542, 0.773 | 165) |
| 169:(27,-37 | 10.768 | 165) |
| 170:(27,-40 | 10.768 | 165) |
| 171:(27,41 | 11.025, 0.256 | 165) |
| 172:(27,-41 | 10.768 | 165) |
| 173:(30,-36 | 10.768 | 167) |
| 174:(30,-37 | 11.542 | 168) |
| **175:(30,40** | 11.762, 0.220 | 168) |
| 176:(30,-40 | 11.542 | 168) |
| 177:(30,-41 | 11.542 | 168) |
| 178:(31,-36 | 10.768 | 173) |
| 179:(31,37 | 11.125, 0.355 | 173) |
| 180:(31,-37 | 11.542 | 174) |
| 181:(31,-40 | 11.762 | 175) |
| **182:(31,41** | 11.986, 0.224 | 175) |
| 183:(31,-41 | 11.762 | 175) |

*Linked list depiction of Viterbi **P**array, with Right and Left half-edge indices, Totalscore and Localscore measures, and Predecessor links. The solution is shown in boldface, starting at* **182** *(read up from the bottom right).*

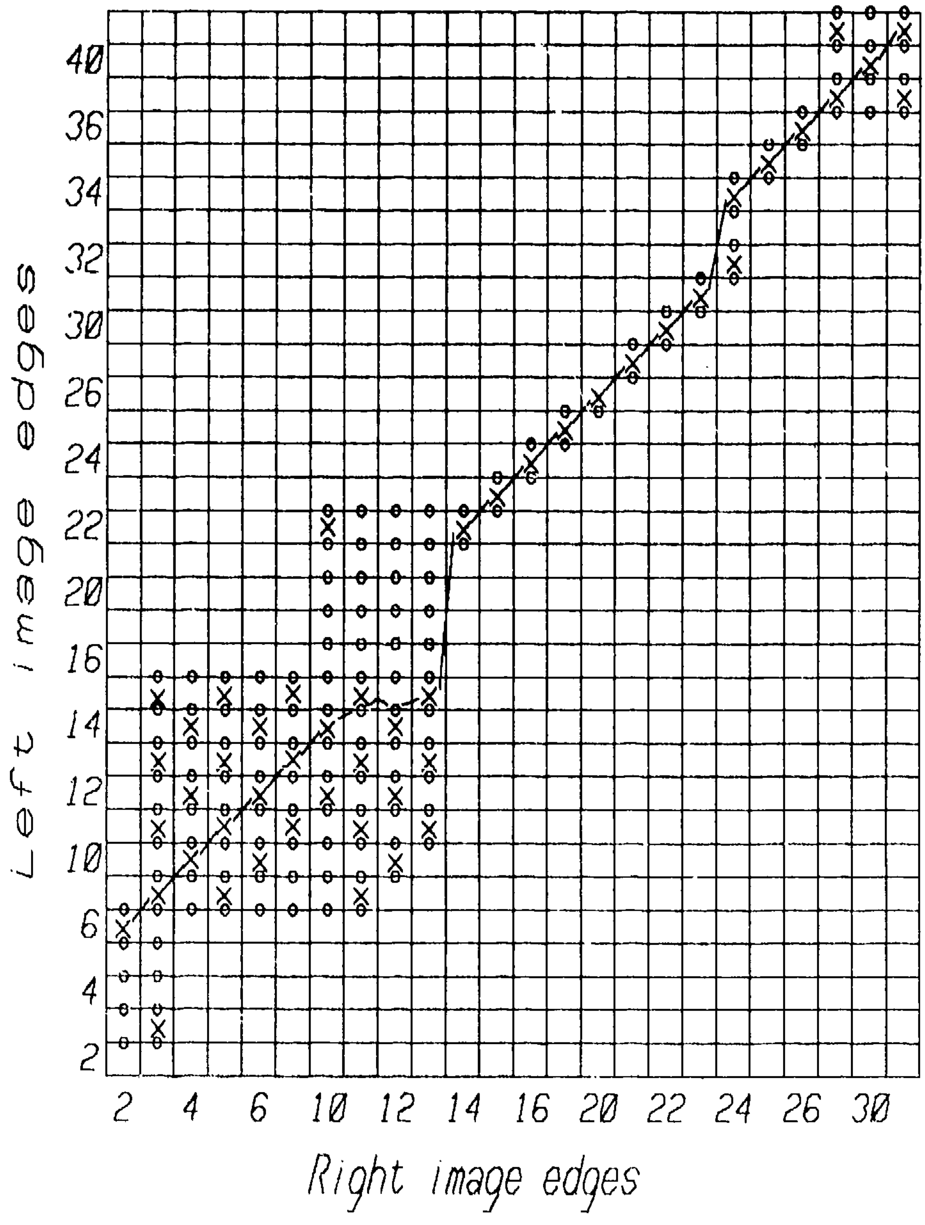
Figure 5-5

In the notation of (5-4), the solution for this line pair is:

$$\{ (2,6),(3,7),(4,10),(5,11),(6,12),(7,13),(10,14),(13,15),(14,22),(15,23),(16,24),$$
$$(17,25),(20,26),(21,27),(22,30),(23,31),(24,34),(25,35),(26,36),(27,37),(30,40),(31,41) \}$$

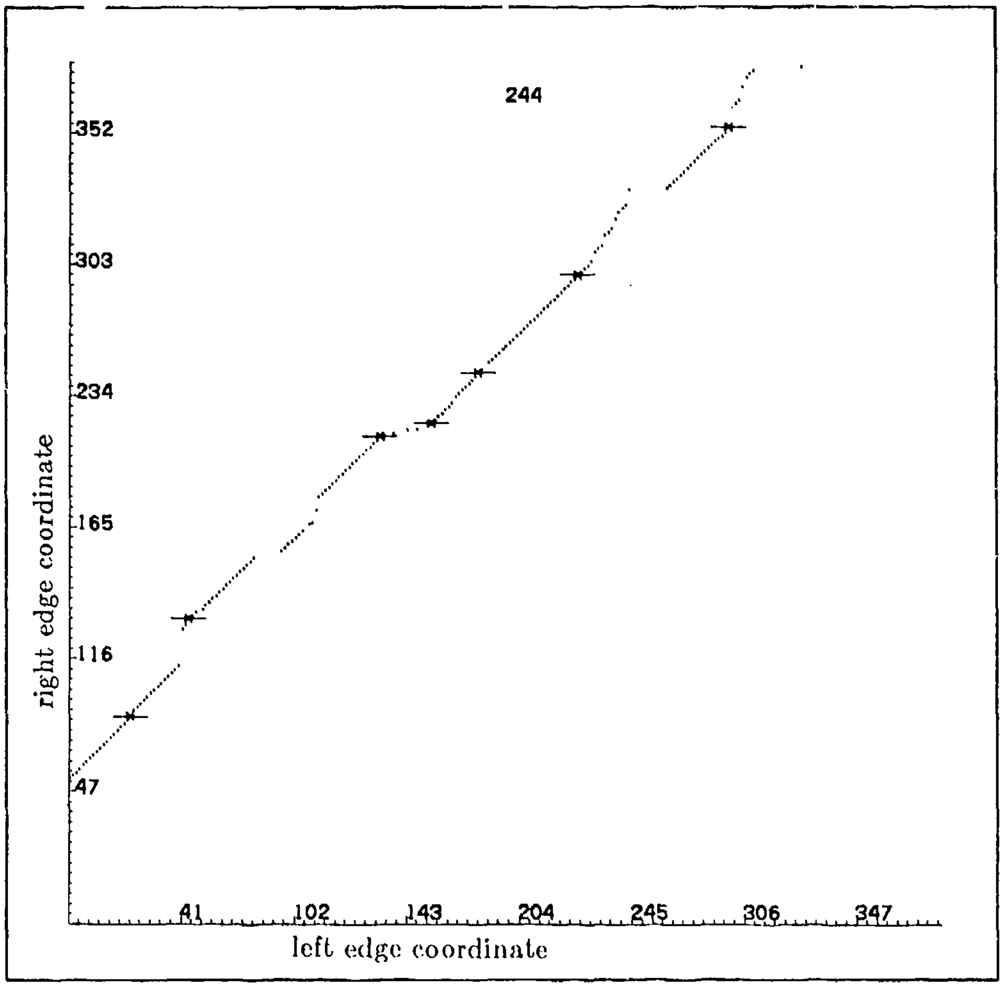
Right image half-edges 11, 12, 32, and 33 have no correlate in the left image.

The consistency enforcement process takes these *locally-based* edge correspondences, removes those which violate global contour continuity (as described earlier), and propagates two-dimensional connectivity in the images to add edge correspondences. This leaves a *kernel* of good correspondences which provide a context for the *constrained-interval edge* and *intensity* correspondence processes. Figures 5-7, 5-8, and 5-9 indicate the result of these final correlations on the line pair shown above (lines $244_8$ in the left and right images). Figure 5-7 plots left image coordinates along the horizontal axis against right image coordinates up the vertical axis. The arrow heads (←→) show the *left* and *right* half-edge *lockings* (of Figures 4-8 through 4-10). The >| and |< symbols indicate edge pairings



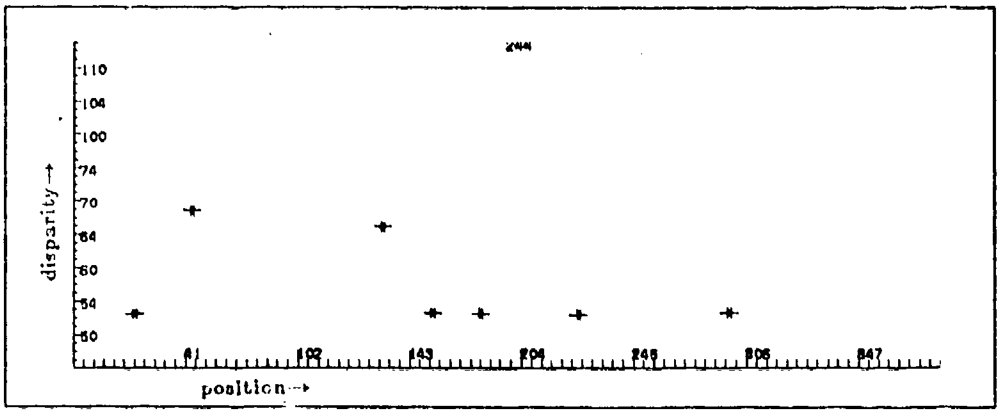
Viterbi array correspondences of left and right image line 244
The x's indicate edge pairings, while the circles show null correspondences
Figure 5-6

added by the *constrained-interval edge* matching, again, with the direction of the arrow indicating the polarity of the *locking*. The little dots mark pixel correspondences. Figure 5-8 is a left image *coordinate* versus *disparity* representation of Figure 5-7. There are unassigned left image edges at those positions where the correspondence process determines the right image pixels are either occluded or are too dissimilar to be matched. Figure 5-9 has an interpolated depiction of the edge and intensity matches of Figure 5-8. Remember that this plot is ***disparity*** — not depth. The two spikes to the right of the figure are the leading edge of the large building at the bottom of the image and a vertical surface of the small building to the right. Since the images are perspective projections, the vertical surfaces (and places of intersection of vertical surfaces — corners) will appear as slanted in a ***disparity*** versus ***position*** depiction. Referring back to Figure 5-4 will clarify this notion.

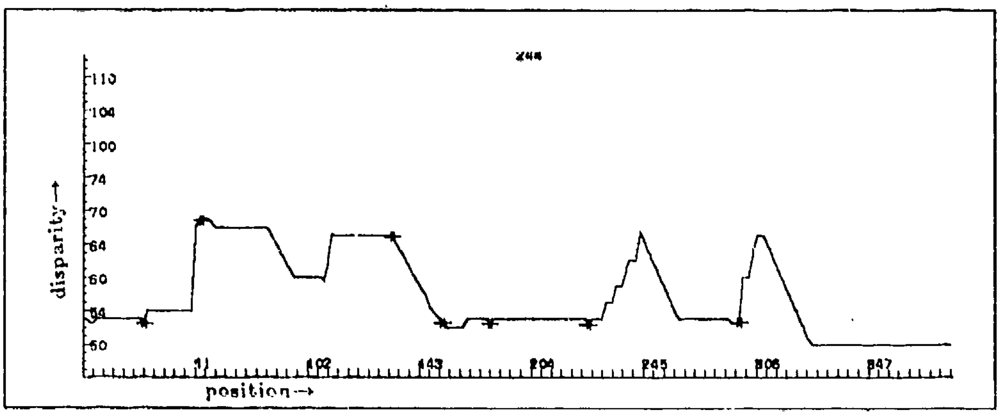


*Edge and Intensity correspondences* (←→ *preliminary edge matches*; |< >| *subsequent edge matches*)

Figure 5-7



*Edge Disparities*
Figure 5-8



*Interpolated Edge and Intensity Disparities*
Refer to Figure 5-4. Notice that the intensity-based matching correctly maps the hollow center of the building to the left, and follows up the edges of the two buildings to the right. The fact that the imaging is a perspective projection makes the building corners appear to be non-vertical (see Figure 5-4) — in fact the vertical vanishing point is at the center of the image. The slope of the wall between the third and fourth double arrows is consistent with the local edge matchings and intensity values — the situation is as suggested in Figure 4-8. Such mappings will be seen to occur fairly often in the intensity-based correspondence process, and show the need for more global analysis.
Figure 5-9

# Chapter 6

# 3-SPACE CONSISTENCY

## 6.1 Using Continuity of Bounding Contours

The edge-based matching described in the preceding sections dealt with line-pairs from the left and right images one at a time. I purposefully kept the analysis from incorporating the results of prior line-pair analyses into the analysis of subsequent line-pairs. Quite obviously there is a strong relationship between the edges on adjacent image lines, and the results of the correlation of one line-pair should be expected to bear some resemblance to the results of the correlation of its adjacent line-pairs. This follows directly from the continuous nature of surfaces. By far the greatest area in our field of view is made up of smoothly varying continuous surfaces — the discontinuities between surfaces occupy only a small (but very important) part of that view. The surfaces are generally continuous, and we expect the bounding contours of those surfaces to be generally continuous.

The edge-based description aims its analysis at those bounding contours — be they boundaries in the intensity domain, as delimit surface detail, or in 3-space, arising as occluding (perhaps self-occluding) contours. The projective connectivity analysis, that part of the edge finding operation which links together neighbouring edges, joins edges that lie along such bounding contours (see Figure 3-8 for a depiction of edge connectivity). One would hope that the correspondence process would assign similar disparity measures to adjacent edges along these contours — if the contour were flat and orthogonal to the line-of-sight then the disparities should all be roughly the same, if the contour were sloping off away from the imaging plane then the disparities of the receding edges should be monotonically decreasing. This relationship of proximal edges having similar disparity can be used as a global constraint on the correspondence analysis.

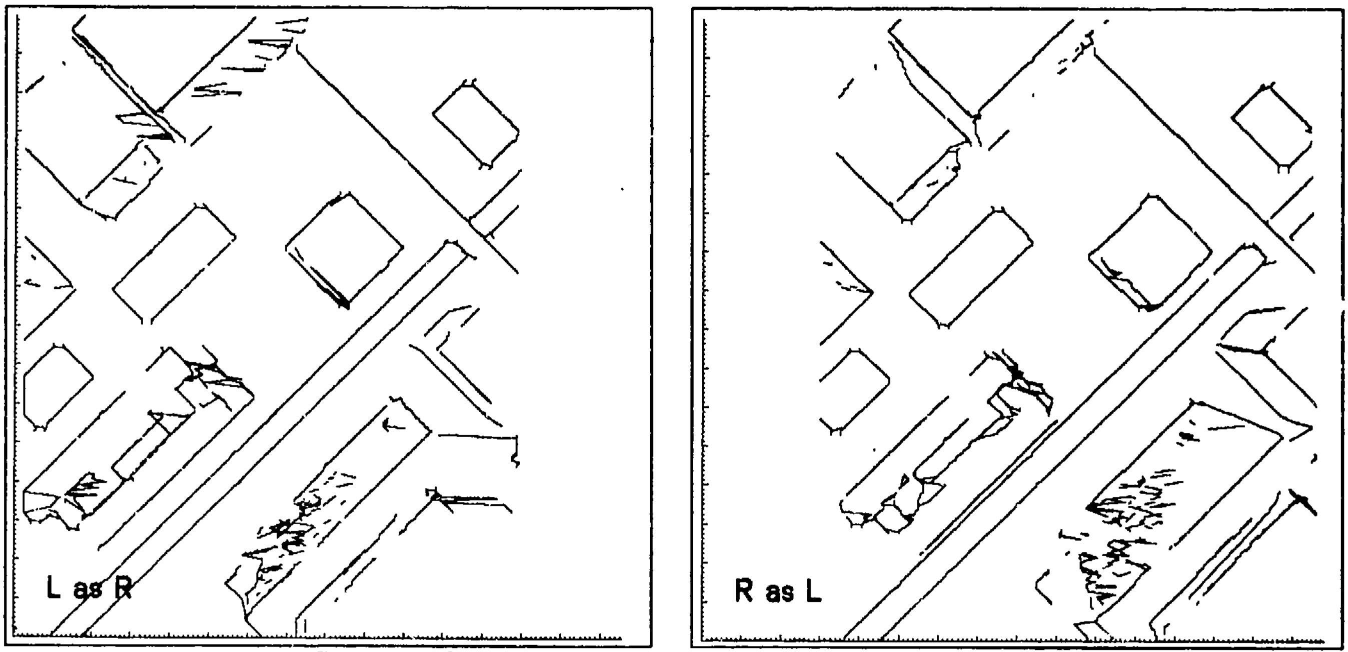
Figure 6-1 depicts these adjacent disparities along connected stretches of edges in the left and right images. In this depiction the connectivity (seen in Figure 3-8) is used to progress from edge to neighbouring edge, but rather than drawing at the coordinates being followed, as in Figure 3-8, the coordinate of the *correlate* of the edge is used. (An alternate way to view this is as drawing the coordinate plus its disparity.) Edges adjacent in the images will be seen nicely connected if they have similar disparities, but will be wildly separated (horizontally) if their disparities differ significantly. Chapter 7 gives a fuller explanation of this depiction technique.

The relationship between continuity in three-space and connectivity in image space is apparent in this depiction. Wherever there is a horizontal deviation between image lines, there is either an abrupt break in contour continuity or, more likely, an error in edge correspondence. So edge connectivity provides inference on the global constraint of contour continuity. The two questions of interest here are — at what stage of the processing should this constraint be introduced, and how should it be implemented in the system?

### 6.1.1 - The introduction of the connectivity constraint

If one were to propagate the results of analysis of line-pair 1 to the processing of line-pair 2, and then these results to line-pair 3, etc., we would be:

- introducing a directional bias to the processing — how would the whole image analysis differ were the processing to run instead from bottom to top?;
- running the risk of sending the correspondence process off into irrecoverable error — the evidence for the matching of edges on certain image line-pairs can be both ambiguous and highly misleading. To make a *single* choice at each line correlation is clearly wrong;



*Correspondence results after local line-by-line processing*
Figure 6-1

- precluding ourselves, from a parallel realization of the correspondence mechanism — the last line would have to wait until all preceding lines were processed.

These options aren't very inviting. If we wish to include the global edge connectivity constraint with the line-by-line analysis, then we are left with only one satisfactory solution — a three-dimensional matching of edges (in the sense that the approach used here is two-dimensional) on *left image edges*, *right image edges*, and *lines*. This is not (as determined yet) an impossible job ... just incredibly complicated and space and time consuming. It is not obvious what the monotonicity constraint would be for the third parameter (lines), nor is it clear that the computation could be ordered so as to be implementable in parallel while maintaining *optimality* (or even be partitionable). This approach deserves future consideration, but is not dealt with further here (see [Moore 1979] for a brief description of a higher dimensional dynamic programming algorithm).

The problems of incorporating the global with the local analysis make it clear (with the above proviso) that the processing of line-pairs should occur independently. But how should we proceed in using the global edge connectivity information?

[Arnold 1982] has devised a scheme for recovering sub-optimal solutions for the individual line-pair correlations, and makes these alternate pairings available to a subsequent consensus forming process. If one were to do a global optimization of all of these pairing possibilities, then this would be a valid approach. However his analysis is local to particular connected stretches of edges.

Another suggestion is to group edges together into *extended edges* or *lines*, making contour continuity explicit. However, the general matching of *extended edges*, which may be fragmented, occluded, etc., is a problem equivalent to the matching of these locally defined edges, so can't be thought of as a fundamental alternative. One of the main points of matching edges, as opposed to larger elements, is the redundancy of information available at this level, and the greater noise-immunity and robustness this brings. Consideration of extended edges can be thought as monocular cueing for the stereopsis, and in this sense would be complementary to local edge analysis.

The philosophy throughout the processing discussed here is, as has been stated before, to work from more reliable signal to less reliable signal, using the results of the higher reliability analyses to guide and constrain the less reliable ones. This attitude sets the direction for the interaction of the line-by-line processings. The role of the first edge-based correspondence process is to provide, if you will, edge-to-edge *locking* between the two images. These *edge lockings* will constrain the registration of the two images. What is being sought is a ***rough global matching of the two images*** — it is not demanded that it be perfect or complete. The flexibility of this target indicates that it would be sufficient for the processing to seek a mutually-consistent *kernel* of edge correspondences. This is done by:

a) allowing line-pairs to be correlated independently, each forming its own assessment of edge correspondences, and then

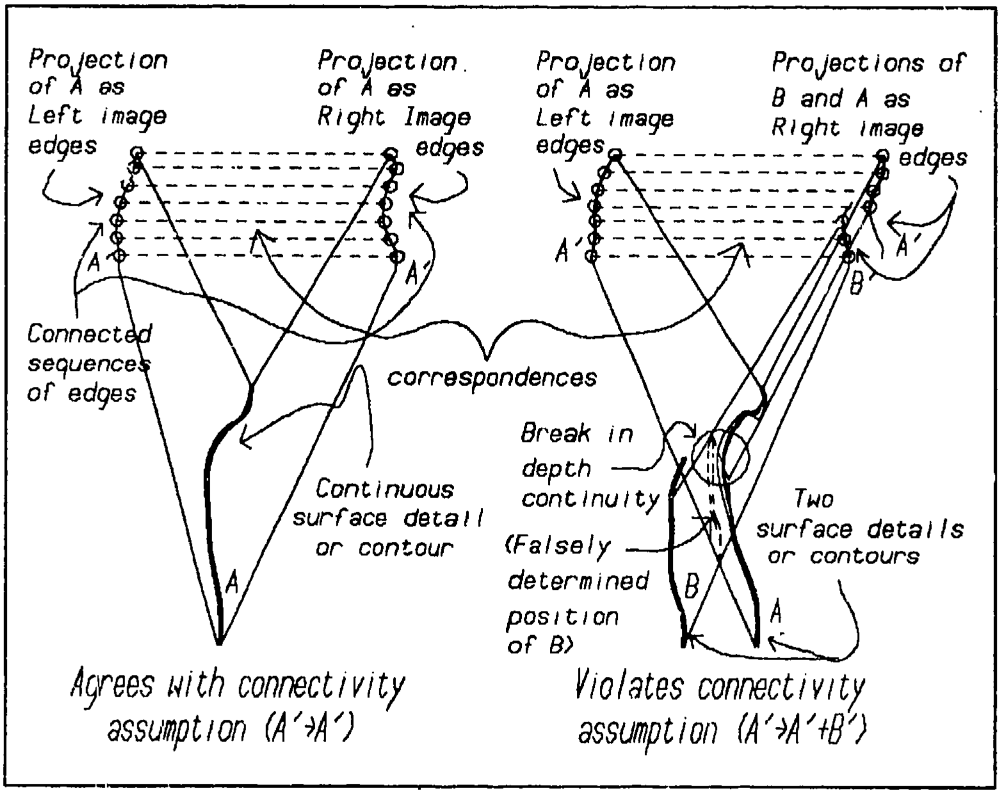
b) cooperatively removing all those correspondences which violate contour continuity.

The next section will describe the implementation of a process to remove globally inconsistent edge correspondences.

### *6.1.2 – The use of the connectivity constraint*

Consider Figure 4-11. Each edge pairing $(R_j \overset{c}{\mapsto} L_i)$ (shown as horizontal lines) has associated with it a disparity $Disp_{i,j}$. The difference in disparity between connected edges (***connectivity*** is shown as vertical lines) is a measure of the implied change in depth between the 3-space points represented by the two pairs of edges. A change in disparity between connected edges that is above some reasonable value will indicate a ***break*** in depth continuity. Except when seen from some anomalous or coincidental viewpoint, a series of edges connected in one of the images will correspond to a continuous bounding contour in the scene. So if there is 2-D connectivity between a series of edges in one image then we should expect their disparities to be smoothly varying. Figure 6-2 illustrates the case of connected correlates along a stretch of edges in both images, and the case of disconnected correlates (which violate 3-space continuity). A measure of smoothness could be obtained by computing the statistics of this disparity first difference distribution. This would yield an interval $[\mu - \sigma_{Disp}, \mu + \sigma_{Disp}]$ of acceptable disparity differences, where $\mu$ is the mean of the computed first differences, and $\sigma_{Disp}$ is the standard deviation.

Further thought suggests that these statistics are actually not appropriate. If there are lots of incorrect correspondences, then the $[\mu{-}\sigma_{Disp}, \mu{+}\sigma_{Disp}]$ interval will tend to be large. Is it reasonable to allow disparity differences over adjacent lines greater than 1.0 (the limit of edge connectivity) to survive? Not likely, as these indicate that the edges in one of the images cannot connect. Failing to have simultaneous connectivity in both of the images is not necessarily a bad sign, though, as gaps frequently occur along any edge path of a contour. The objection here is with edges that are on adjacent lines and **can't** be connected because of their relative horizontal displacement. Would it be reasonable to reject correspondences giving rise to disparity differences that are within the computed interval but outside of the connectivity range? Yes, if they are unrealizable. What about the opposite situation, where the standard deviation $\sigma_{Disp}$ is less than 1 pixel width, with most correspondences good. Here the interval limit will suggest excluding pairings of disparity difference somewhat less than 1.0? Well, after a little thought it becomes clear that the 3-space consistency process should not use these statistical measures, but rather employ a simple distance measure which would prevent discontinuous edges from being matched to the same structure. If the horizontal separation of edges on adjacent image lines exceeds this measure (chosen to be a single pixel width, $\sigma_{Disp} = 1.0$), then the two edges cannot be joined in depth. It is important to note that this puts a limit on the inclination of edges to the line of sight ... connected edges must be discernable as connected within the 2-space imaging resolution for them to be accepted by the matcher as connected in 3-space.


Projection of A as Left image edges
Projection of A as Right Image edges
Projection of A as Left image edges
Projections of B and A as Right image edges
A
A'
A'
A'
B'
Connected sequences of edges
correspondences
Continuous surface detail or contour
A
Break in depth continuity
(Falsely determined position of B)
Two surface details or contours
B
A
Agrees with connectivity assumption (A'→A')
Violates connectivity assumption (A'→A'+B')

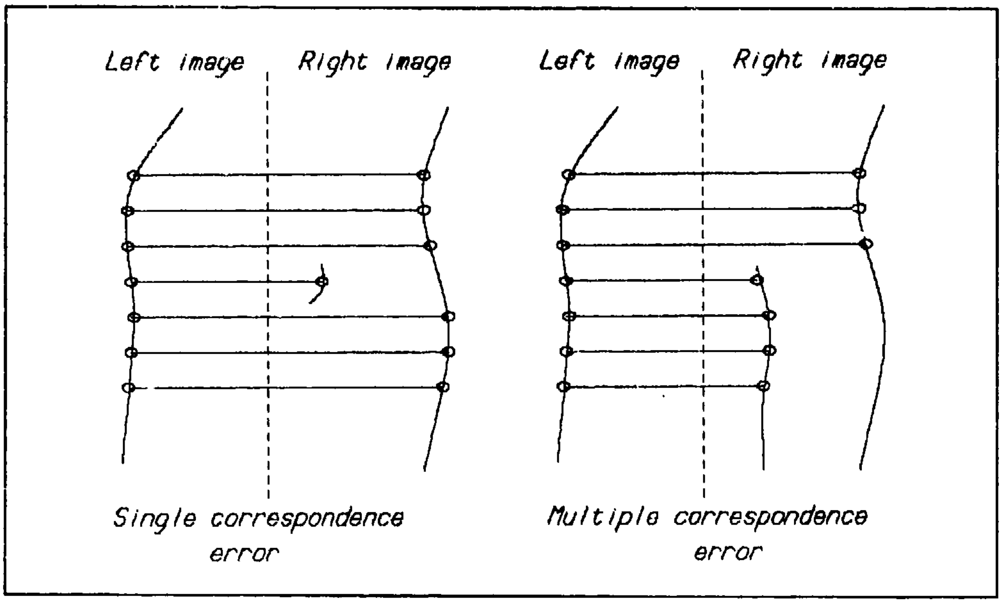
*Two cases of edge matching: left is what is expected, right is a continuity failure.*

Figure 6-2

### *6.1.3 – Implication of large disparity differences*

Identifying the incorrect edge correspondence from an inconsistent disparity difference is not an obvious process. The problem one immediately sees is that a disparity difference outside of the range $[\mu - \sigma_{Disp}, \mu + \sigma_{Disp}]$ conclusively implicates **neither** of the contributing edge pairings. It merely suggests that one of them is inconsistent. Should they both be removed to be sure that the incorrect one is taken out? Not likely, as this conservative policy would lose too many good correspondences in clearing out the bad ones. Consider the case in Figure 6-3 of a single incorrect correspondence bounded above and below by properly assigned correspondences. The two on the periphery could vote, and throw out the offending middle correspondence. Think of this as a single-bit error corrector.

If errors were scattered, like this, rather than systematic, then this simple voting technique would be all that was required. However systematic errors of correspondence occur as well. Consider a case where long stretches of edges in one image are deemed to correspond to some stretch of edges in the other image, then switch *en masse* to correspondence with some other stretch of edges further down the image. Figure 6-3 also depicts an occurence of this situation. The only inconsistent disparity difference here would appear at the junction between the apparently consistently connected stretches. In a worse case situation, a correspondence from the properly associated stretch of edges would be removed for each correspondence incorrectly assigned to the other stretch. A good removal strategy would be one which minimizes the loss of correct edge correspondences.



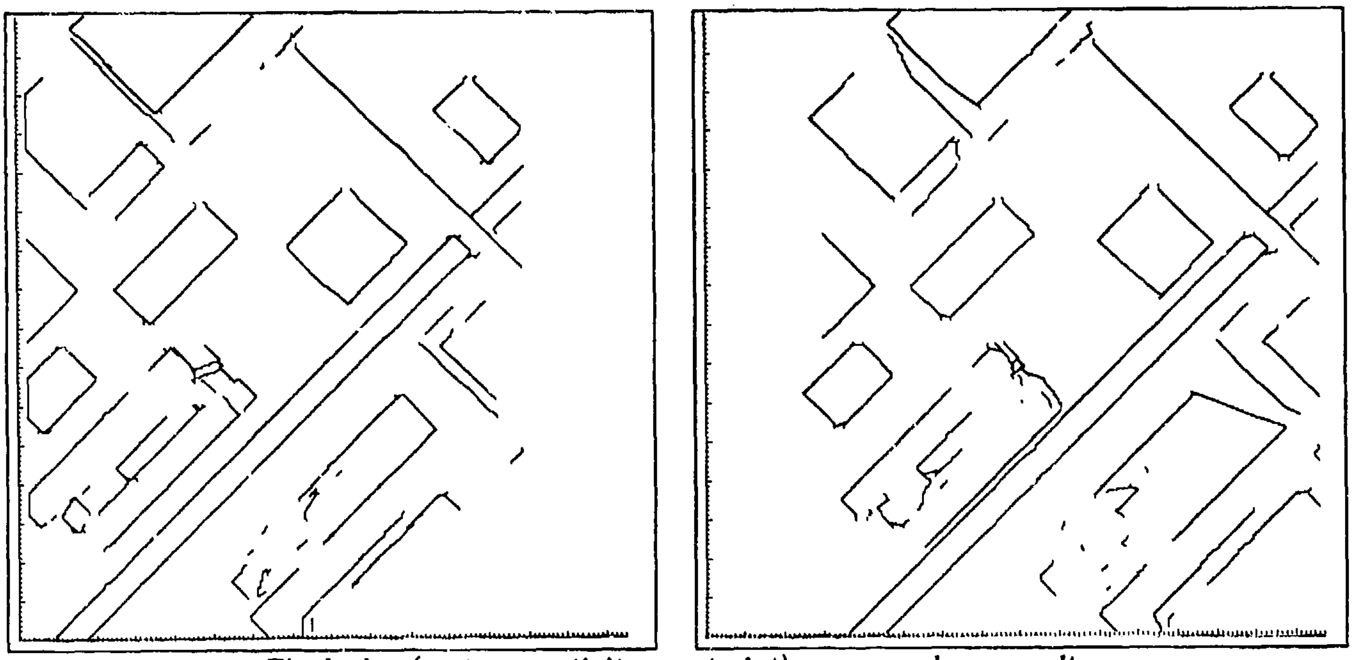
*Single and Multiple Correspondence Errors*
Figure 6-3

## *6.2 Cooperative Connectivity Enforcement Algorithm*

The mechanism developed for this minimal loss strategy is as follows:

1) Flag correspondences incident on an inconsistent disparity difference as ***questionable***.
2) If, in so flagging, the correspondence is found **already** to be flagged, then mark the correspondence for removal.
3) Do 1 (and 2) until all *questionable* correspondences have been flagged.
4) Remove all marked correspondences and re-evaluate the disparity differences about the *newly* connected edges.
5) Do 3 and 4 until no further correspondences can be marked for removal.
6) Remove all flagged correspondences.
7) Do 5 and 6 until no inconsistent correspondences remain.

This algorithm deletes a minimum of valid correspondences, and guarantees the removal of all inconsistent disparity differences. Figure 6-4 shows the connectivity of Figure 6-1 after the inconsistent correspondences have been removed by this cooperative connectivity enforcement algorithm.



*Final edge (post-connectivity constraint) correspondence results*

**3000 half-edge correlate pairs**

Figure 6-4

# *Chapter 7*

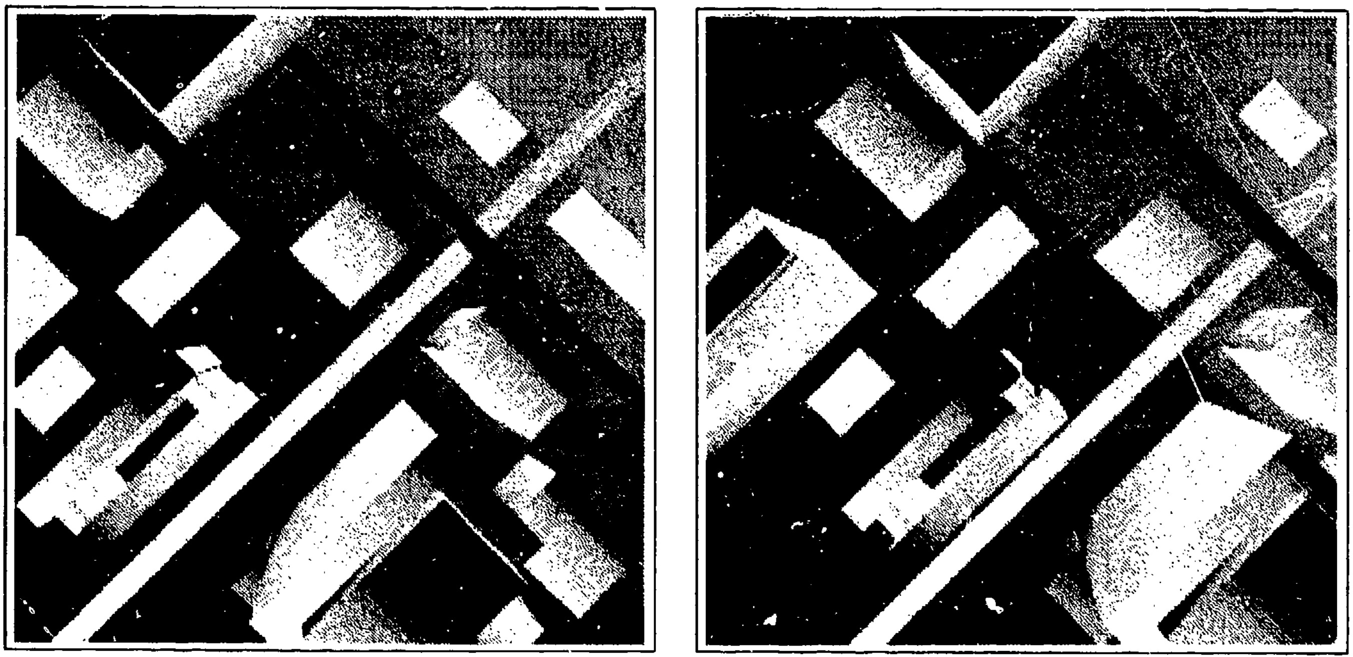
# STEREO CORRELATION OF SAMPLE IMAGERY

The best way to understand the functioning of the total stereo algorithm is no doubt through examples of its processing. This chapter will show you, a step at a time, what is involved in the analysis of some typical imagery and demonstrate how effectively it works.

## *7.1 Control Data Corporation Imagery*

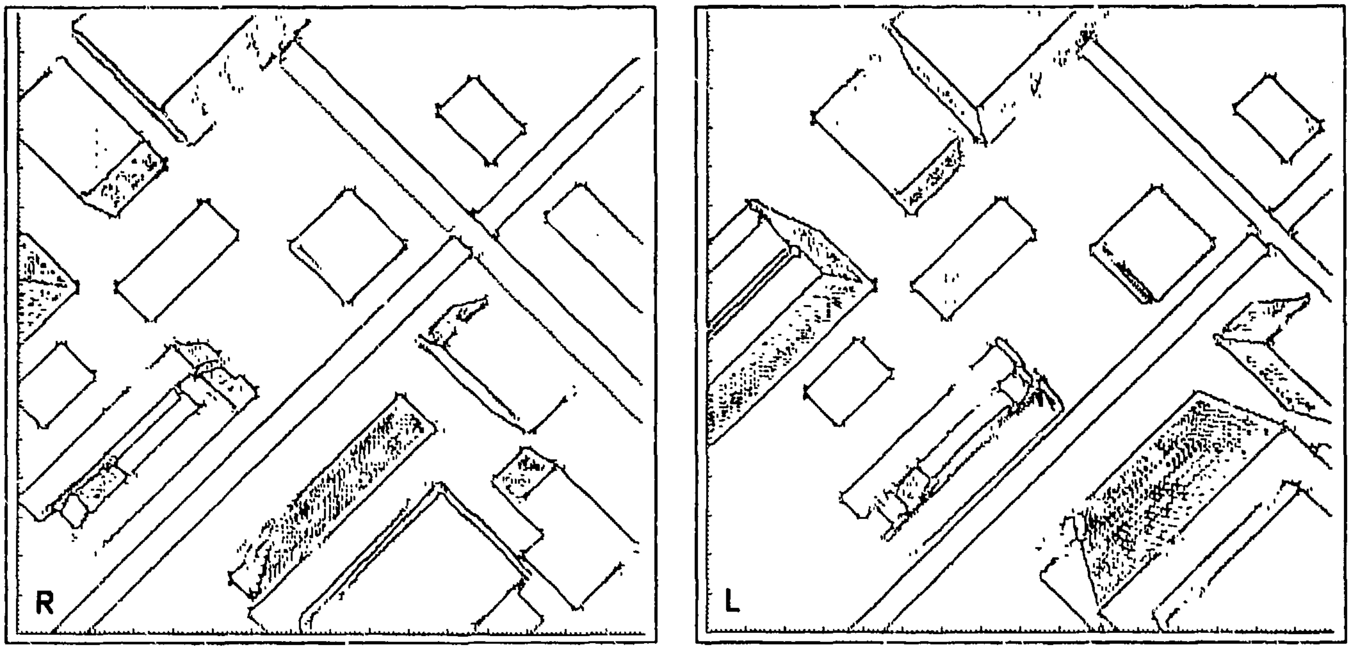
The input to the process is a pair of collinearized stereo images, as shown in Figure 7-1 intensity enhanced. Scan lines in these images correspond to epipolar lines. The stereo pair was created to demonstrate graphics capability rather than to serve as data for a stereo correlator, so exhibits several unappreciated characteristics — it has multiple light sources (making the projections of certain structural edges appear to be discontinuous), and has in effect zero random sensor noise (all noise is from the sampling and quantization).

The standard deviation in intensity variation for this imagery was sampled and estimated as being 0.596, indicating that any first difference above $0.596\sqrt{2} = 0.840$ should be considered to be signal rather than noise. Because of this low noise measure, the reduced resolution matching does not go beyond a single reduction. Figures 7-2 and 7-3 show the full resolution and reduced resolution ($T = 1$) edges found for this imagery. Figure 7-4 shows the connectivity between the various edges of these two images (recall that edge connectivity plays a part in the global consistency analysis). Figure 7-5 is a broadened depiction of the intensities along a pair of corresponding lines of this imagery while Figure 7-6 shows the full and reduced resolution edges found along these lines. The

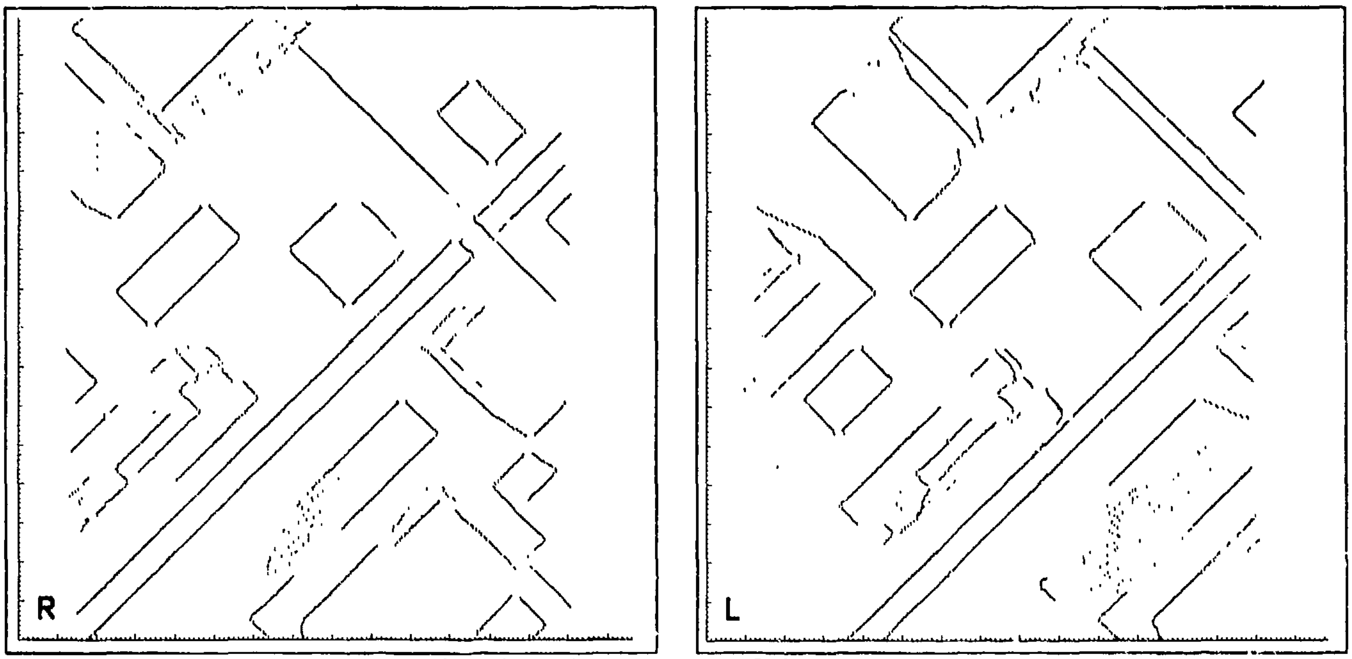


*A stereo pair of images (from Control Data Corporation)* [256 × 256 × 6]
Figure 7-1

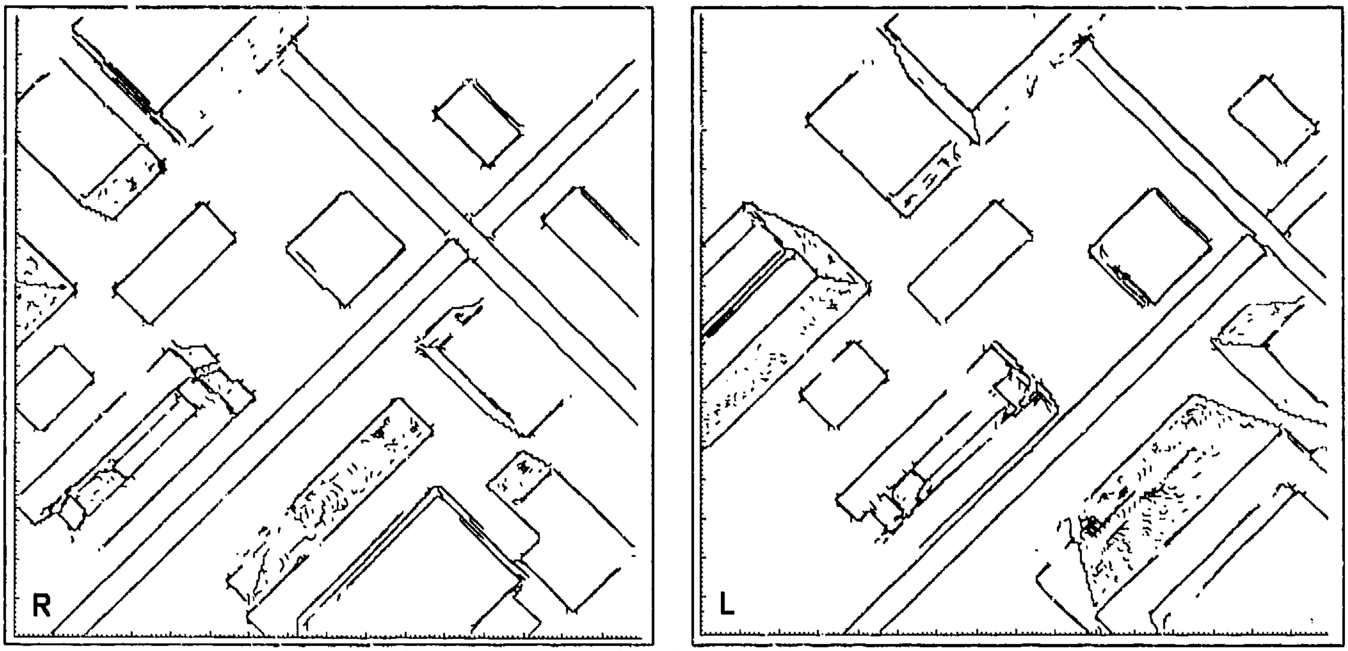
edges — the top and bottom position of the vertical lines indicate the intensity averages to the left and right sides of the edges, and horizontally sloping lines show the interpolated intensity gradients in the intervals between image edges. Diagonal marks in the upper profile of the figure indicate edges paired by the reduced resolution matching.



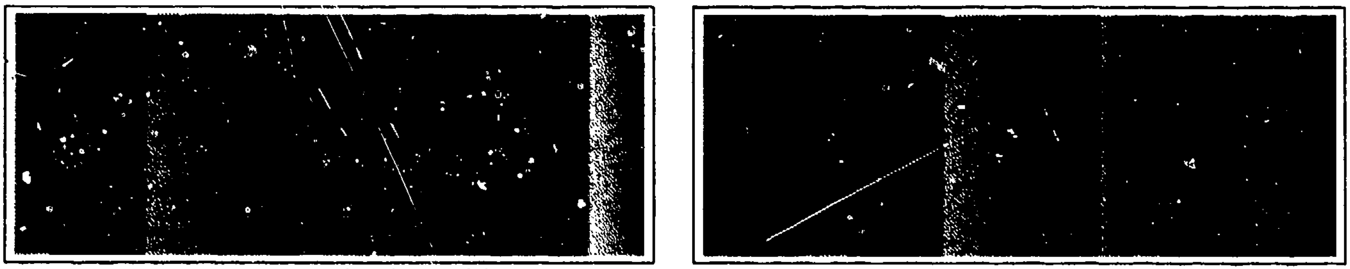
*Full resolution edges of the stereo pair*
Figure 7-2



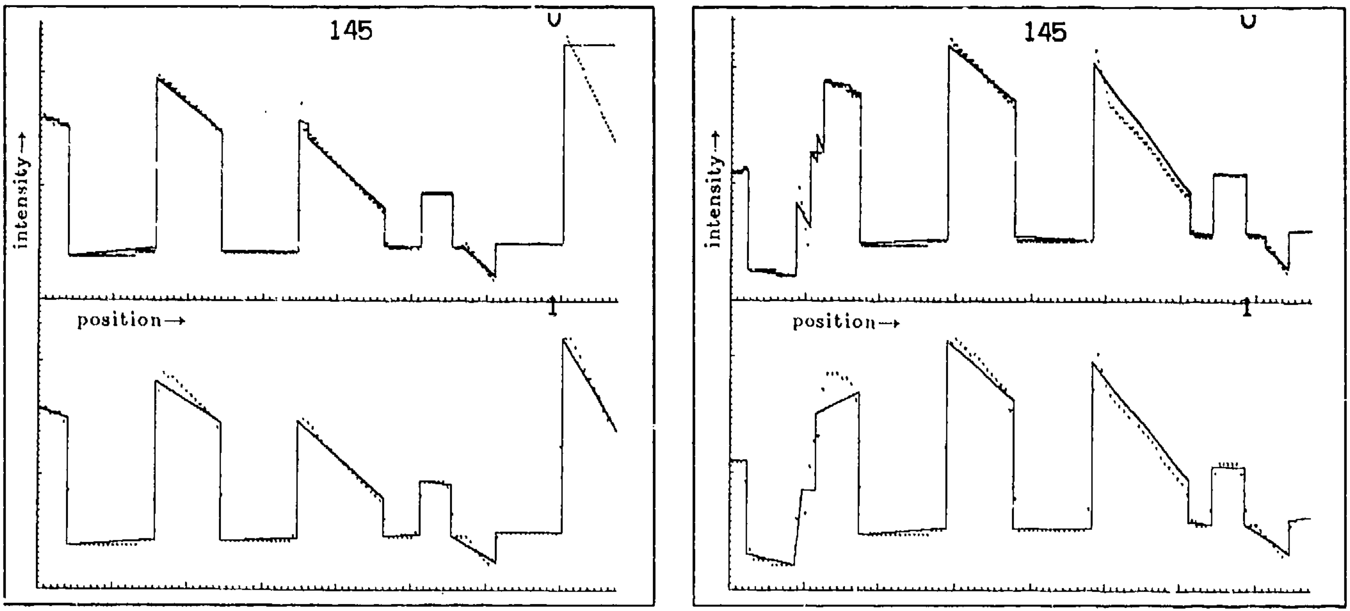
*Reduced resolution edges of the stereo pair*
Figure 7-3



*Connectivity of the edges of the stereo pair*
Figure 7-4



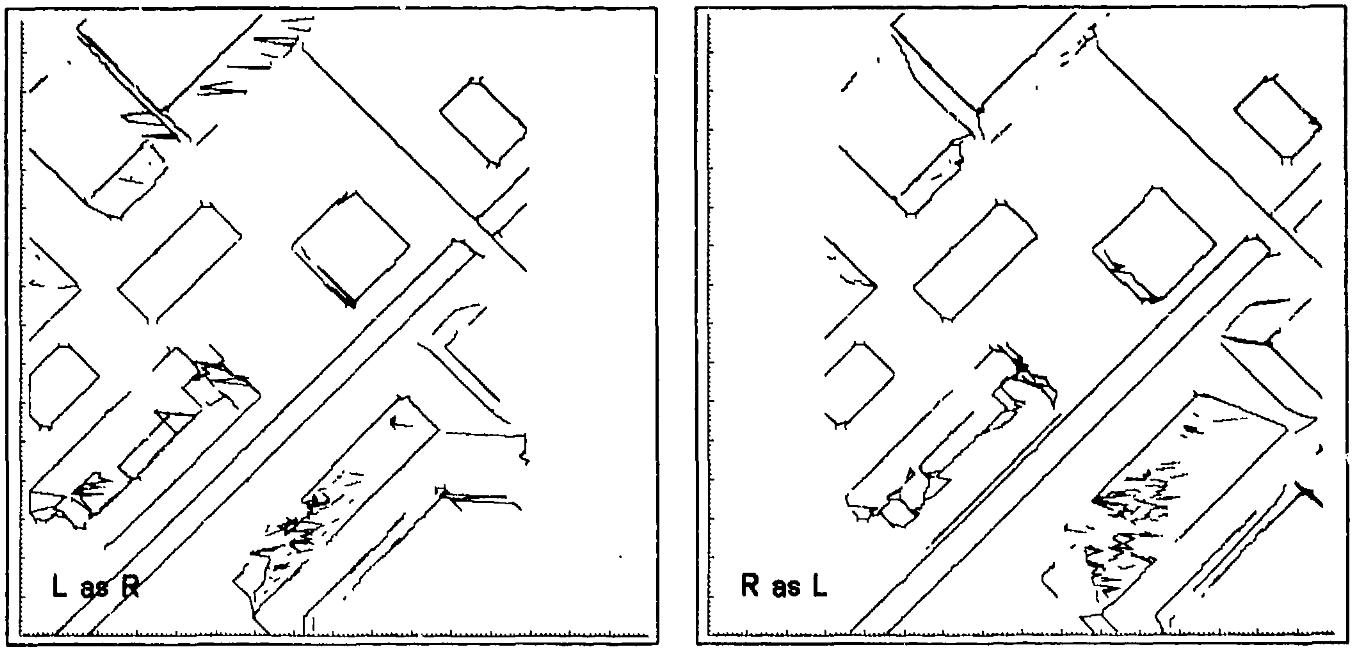
*Right and Left image corresponding line intensities*
Figure 7-5



*Edges of this line-pair at full and reduced resolutions*
Figure 7-6

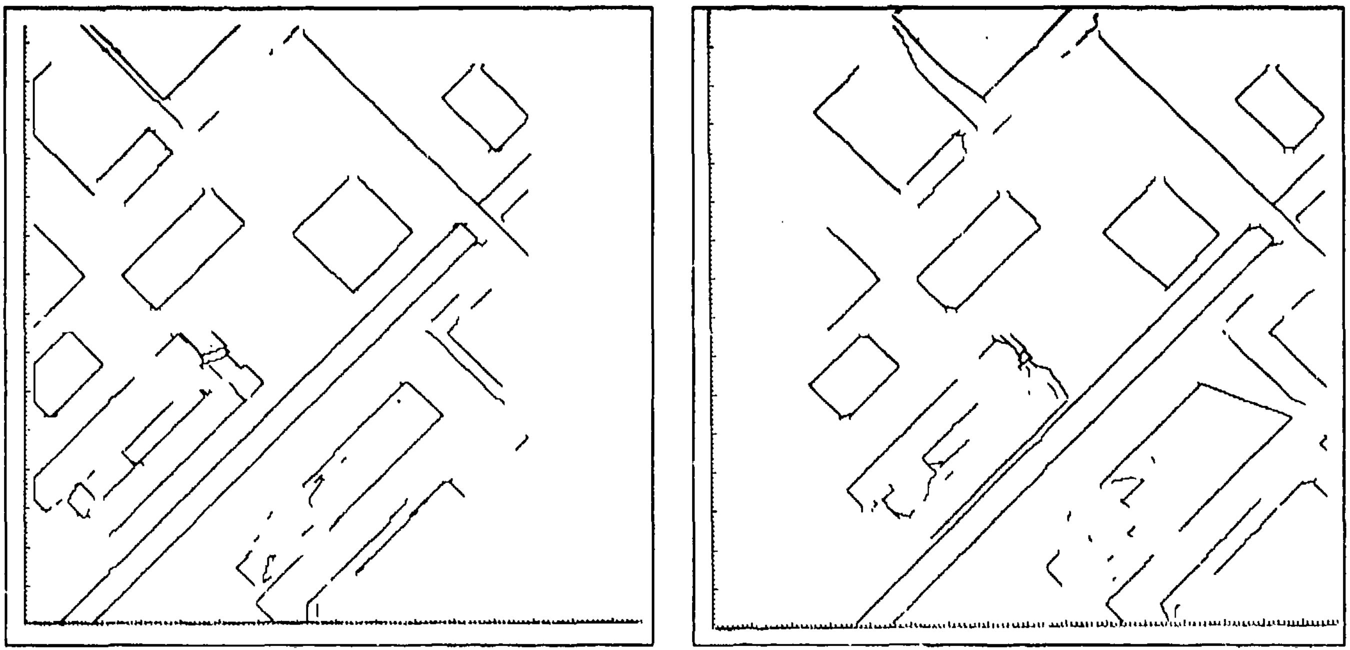
The reduced resolution and full resolution edge matchings process line-pairs such as these, determining the best line-by-line correspondences. Figure 7-7 shows the results of this processing for the CDC imagery. The depiction may be difficult to understand:

- The left figure shows the edges of the left image, drawn with their connectivity (as Figure 7-4 left), but rather than using the coordinate of the left image edges, uses the coordinates of their mates in the right image (this is equivalent to using the coordinate plus associated disparity).
- The right figure shows the edges of the right image, drawn with their connectivity (again, Figure 7-4 right), but rather than using the coordinates of the right image edges, uses the coordinates of their mates in the left image (which is the same as the coordinate minus associated disparity).
- Since the lines joining connected edges are all that are being drawn, if two adjacently connected edges in one image, for example the left, are found to match two unconnected edges in the other image, then the line joining them in the left figure will run (nearly horizontally) as though between the two disparate edge coordinates. What this reveals, and reveals quite clearly, is the correlation's decision that there is a variation in depth between the two matched pairs of edges. In general, horizontal lines suggest errors in the correlation (notice that there are relatively few in this depiction).

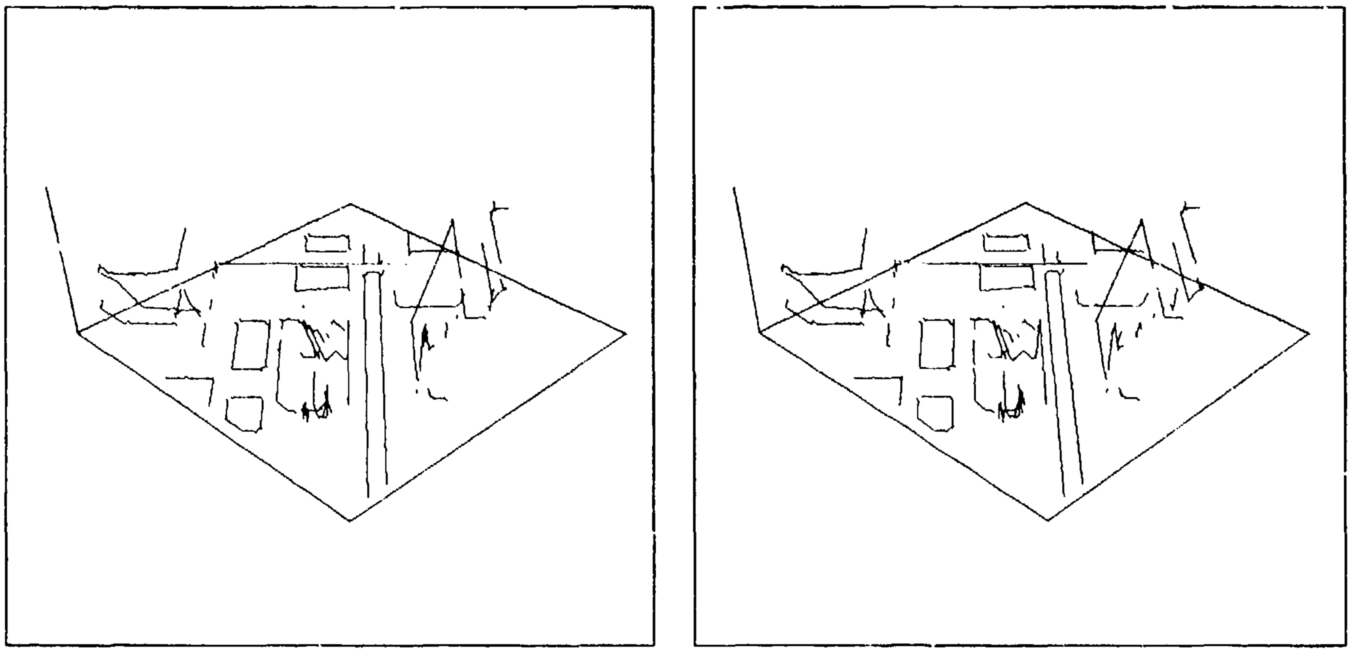


*Preliminary matching results*
Figure 7-7

The cooperative process that ensures global consistency removes inconsistent matches, propagates disparities along connected edge paths, and results in a *kernel* of sound correspondences. These final edge-based matching results are shown in Figure 7-8. The figures are drawn in the manner of Figure 7-7. The stereo depiction of Figure 7-9 is a perspective view of the connectivity shown in Figure 7-8 (which was shown there from directly overhead).

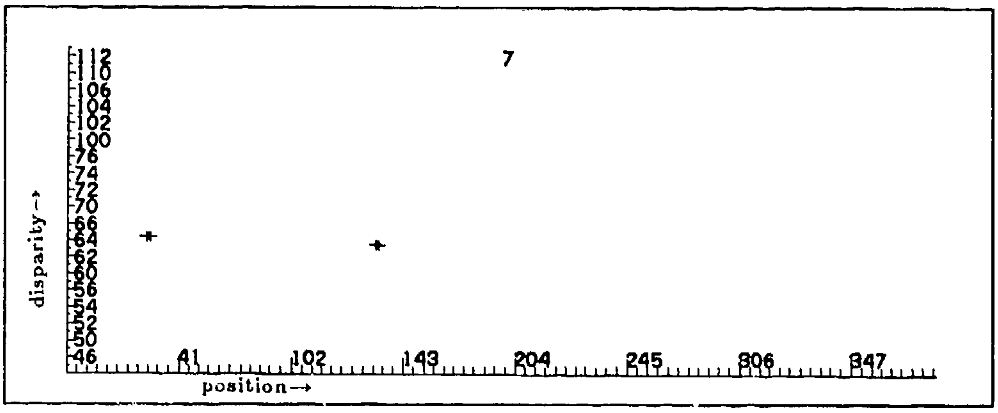


*Final (post-connectivity constraint) matching results*
**3000 half-edge correlate pairs**
Figure 7-8



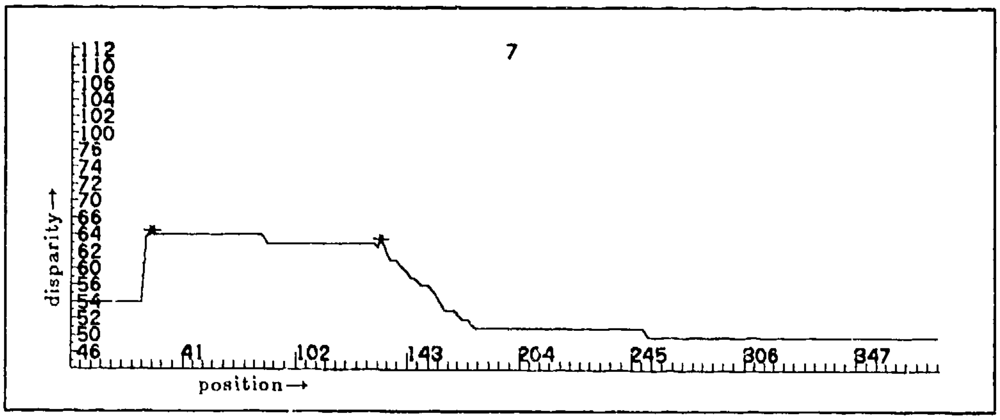
*Perspective view of connected edge elements*
Figure 7-9

In the phrasing used earlier, the matching results at this stage form a ***template of constraints*** for the next stage of the processing. Considering the edge-based correspondences on a line-by-line basis, we can think of the edge matchings as defining a local mapping of intervals between the two images. Edges in the corresponding intervals that have not been assigned matches by the prior correspondence process are candidates for matching within this more tightly constrained context. The processing of an ***interval-constrained edge***-based matching completes edge matching in the intervals, and a **final** correlation, using the intensity values of the pixels themselves, ***interval-constrained intensity*-based** correlation, determines pixel to pixel correspondences. Figures 7-10 and 7-12 show the matching of edges attained through the edge-based correlation for several image line-pairs. Original **edge** correspondences are indicated by arrows → and ←, where the left arrow positions a **right *half-edge*** and the right arrow positions a **left *half-edge***; subsequent ***interval-constrained*** edge **correspondences** are indicated by >|and |<. Individual comments appear on the figures themselves. An **interpolated** disparity representation of these same graphs can be seen in Figures 7-11 and 7-13 (this **display is** perspective, so verticals have varying horizontal components).



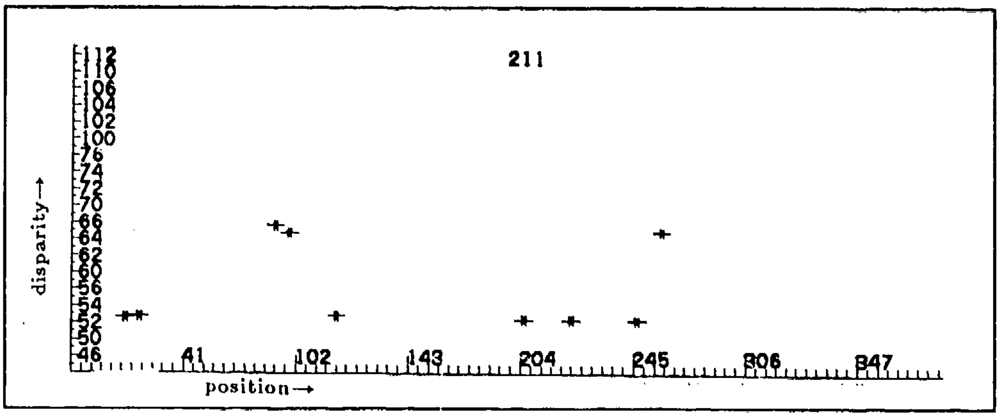
*Edge correspondences* (←→ *preliminary edge matches*; |< >|*subsequent edge matches*)

Figure 7-10

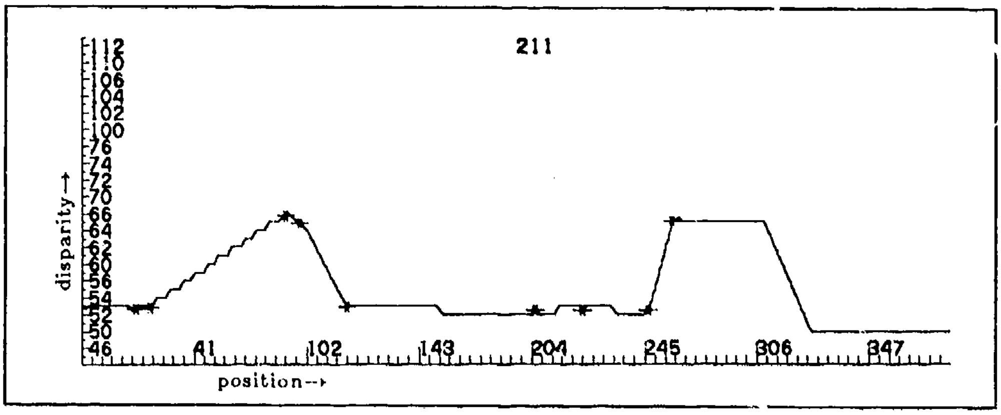


*Interpolated disparities*

Figure 7-11



*Edge correspondences*
Figure 7-12

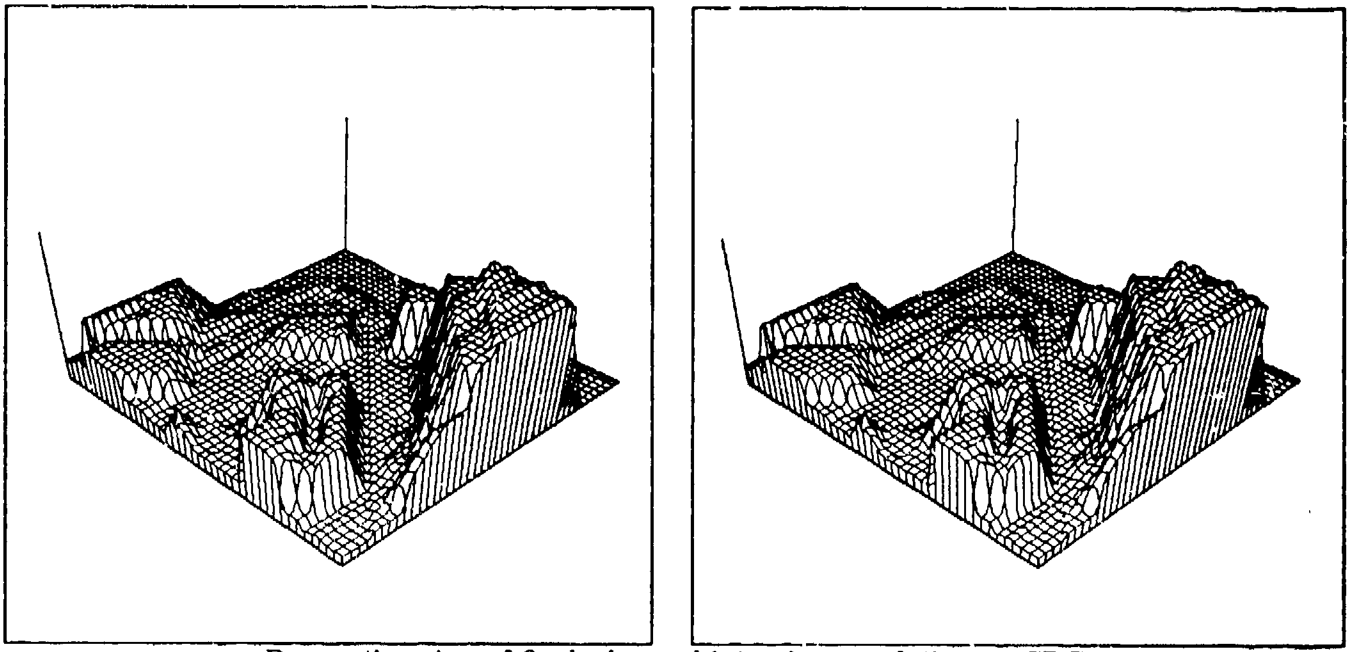


*Interpolated disparities*
The surface slope between the second and third double arrows arises for the same reasons as it did in Figure 5-9.
Figure 7-13

Figure 7-14 shows the full image array disparity map — the result of the processing of the four correlations:

1) reduced resolution edges,
2) full resolution edges,
3) interval-constrained edges,
4) interval-constrained pixels.

The depiction is again perspective, and shown from the point of view of the right CDC image. Without knowing the camera parameters, or at least the relationship between the two sets of camera parameters, there is no possibility of transforming the representation to an orthographic form. I do, however, have an interactive program that allows estimates to be made on the transformation, and this produced the orthographic correction for the perspective stereo plot of Figure 7-15 (which is a half resolution depiction — Figure 7-14 was smoothed and sampled at one third resolution for increased clarity). Figure 7-16 is a monocular depiction of the perspective projection at full resolution.

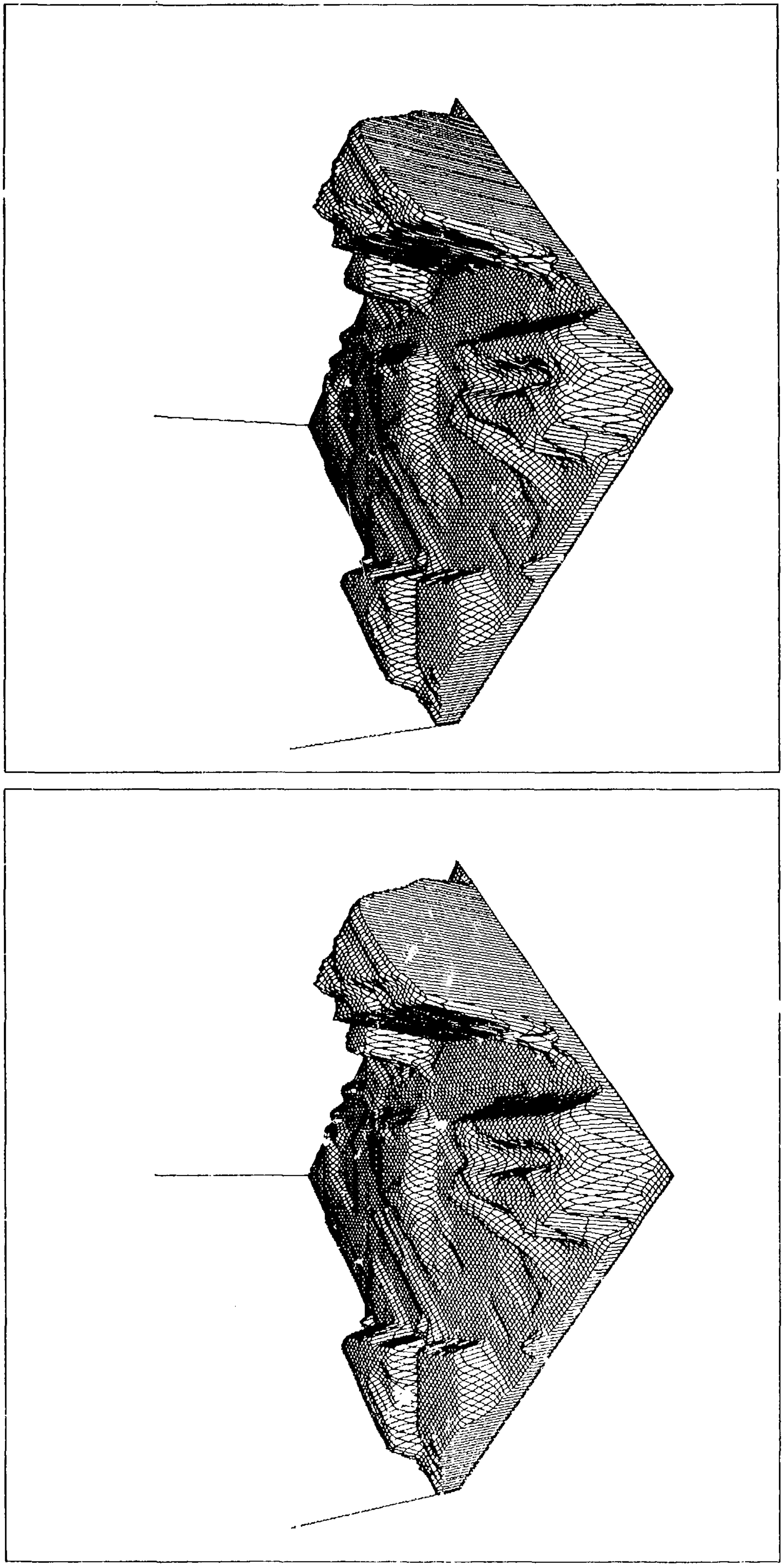


*Perspective view of final edge and intensity correlation — CDC*

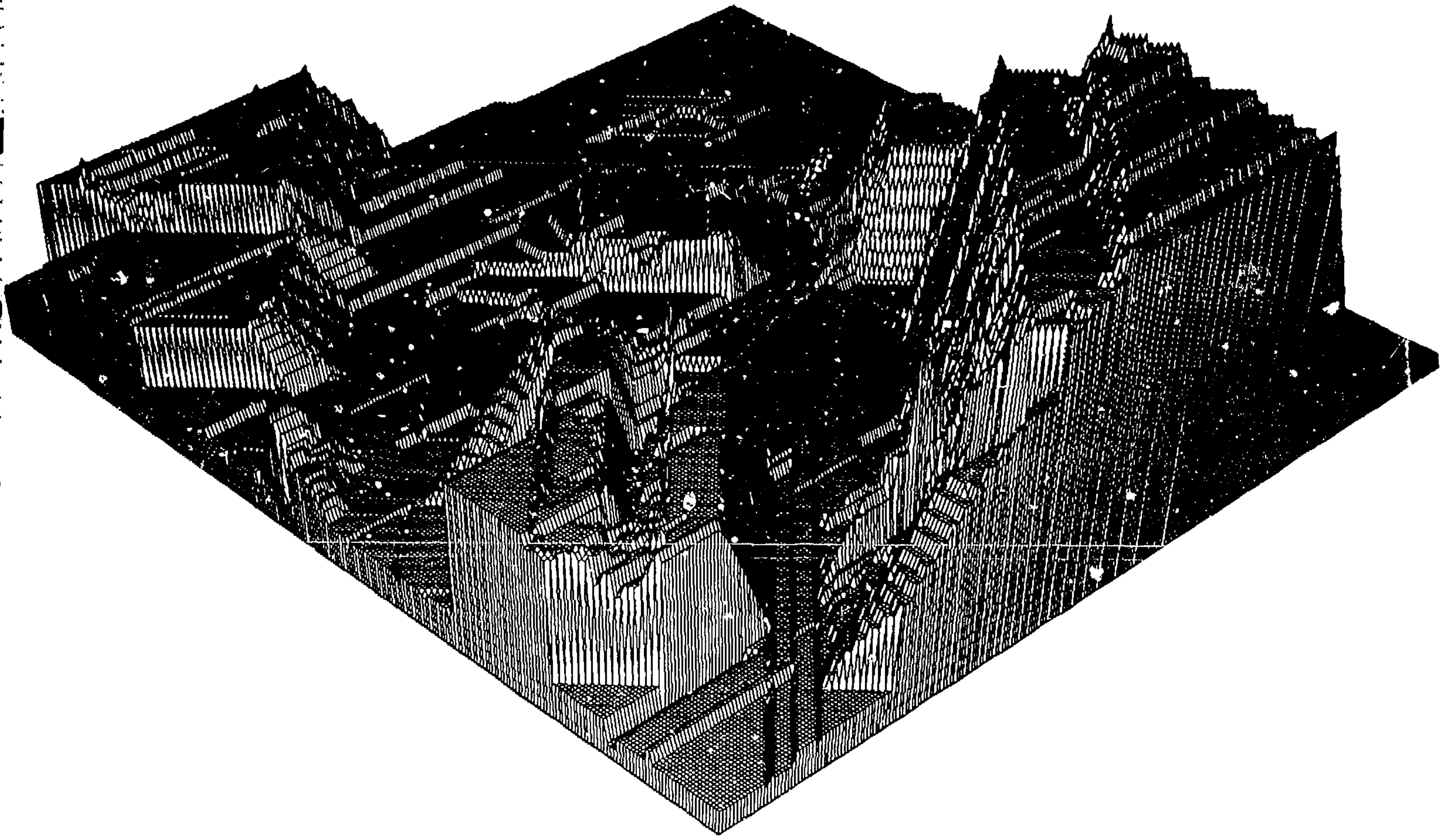
(the $z$ axis is disparity, not elevation)

The left side of the low building in the upper center, and the far left top side of the nearest building (the hollow one) show incorrect surface slopes (as in Figures 5-9 and 7-13). The near left top side of the same hollow building extends too far, running to the edge of the image. Again, the intensities alone do not provide sufficient information for a correct positioning of these surfaces (they should be in the ground plane). More global surface information is available, although unused here, and this will provide better positioning constraints when further refinements are made to the intensity correlation algorithm.

Figure 7-14



*Smoothed and sampled orthogonal depth map – CDC*
(the *z* axis is elevation)
Figure 7-15

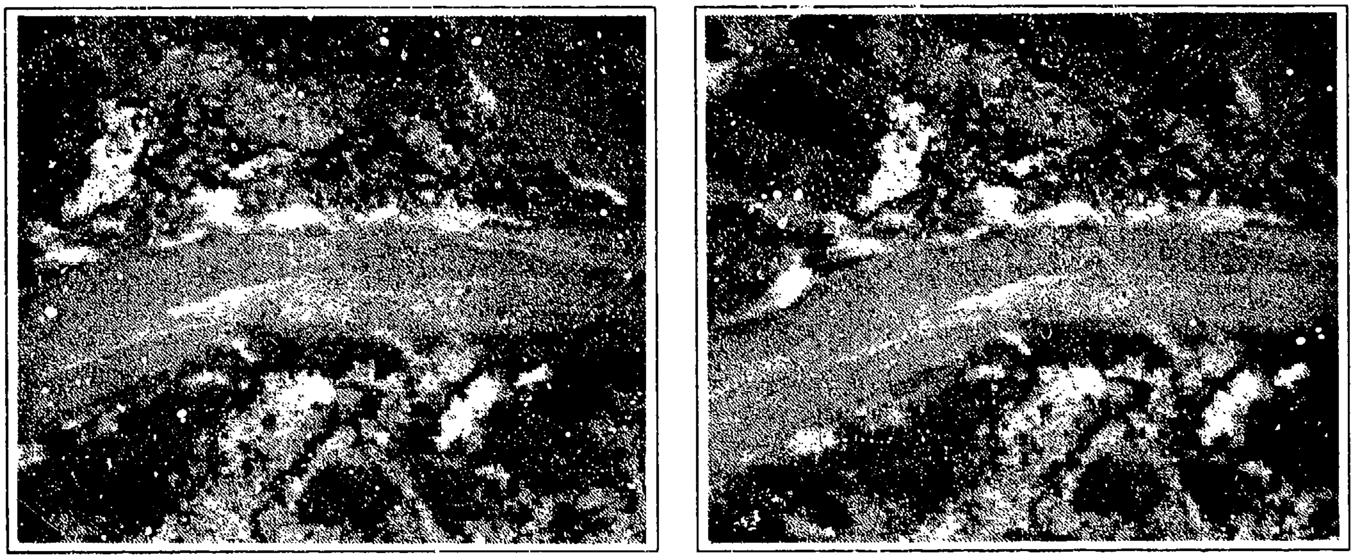


*Full resolution CDC plot (of Figure 7-14) before orthogonalizing*
Figure 7-16

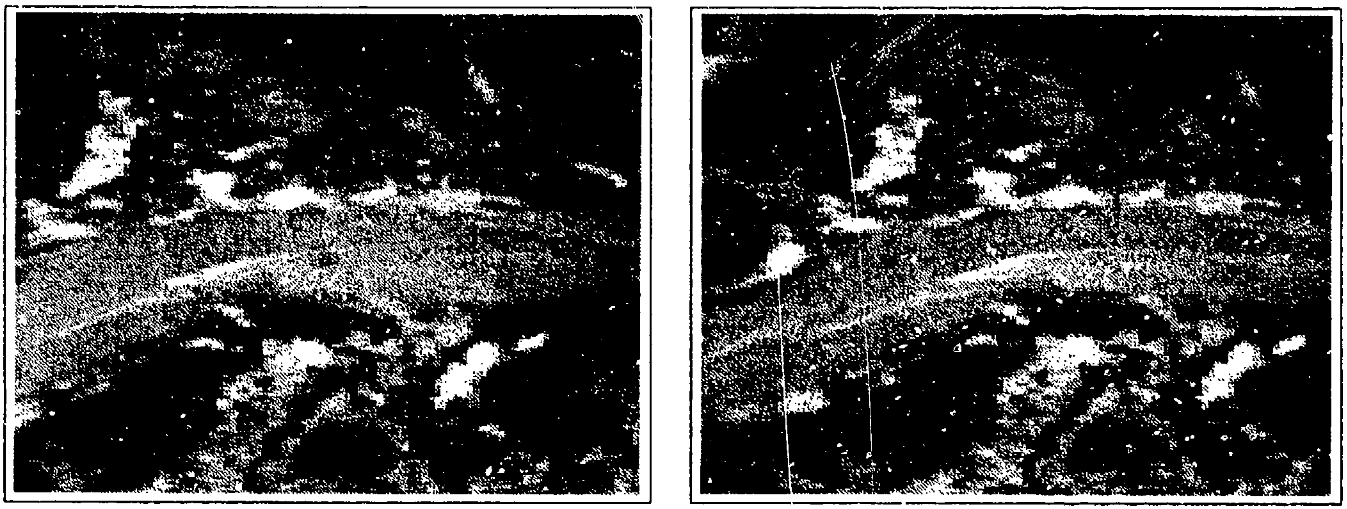
## 7.2 Night Vision Laboratory Imagery

Another example of the stereo processing is shown in Figures 7-17 through 7-35. The imagery depicted here, a valley scene recorded on videotape, was provided by the Night Vision Laboratory of the United States Army. The scene is synthetic in that it is a papier-mâché recreation of an actual valley, although the imaging is real. There is very little relief in the scene, although it has a general drift to higher elevations toward the upper left corner. There is only slight difference in height between the river (running through the centre across the images) and the various land and vegetation areas. Figure 7-17 shows the stereo pair at full resolution, while Figures 7-18, 7-19 and 7-20 show the three resolution reductions (reduced to the limit for noise suppression). Figure 7-21 shows the edges determined for the full resolution image, and Figure 7-22 shows the edges determined by the largest convolution operator for the most reduced resolution image (Figure 7-20). Figure 7-23 depicts the edge connectivity for the full resolution images.

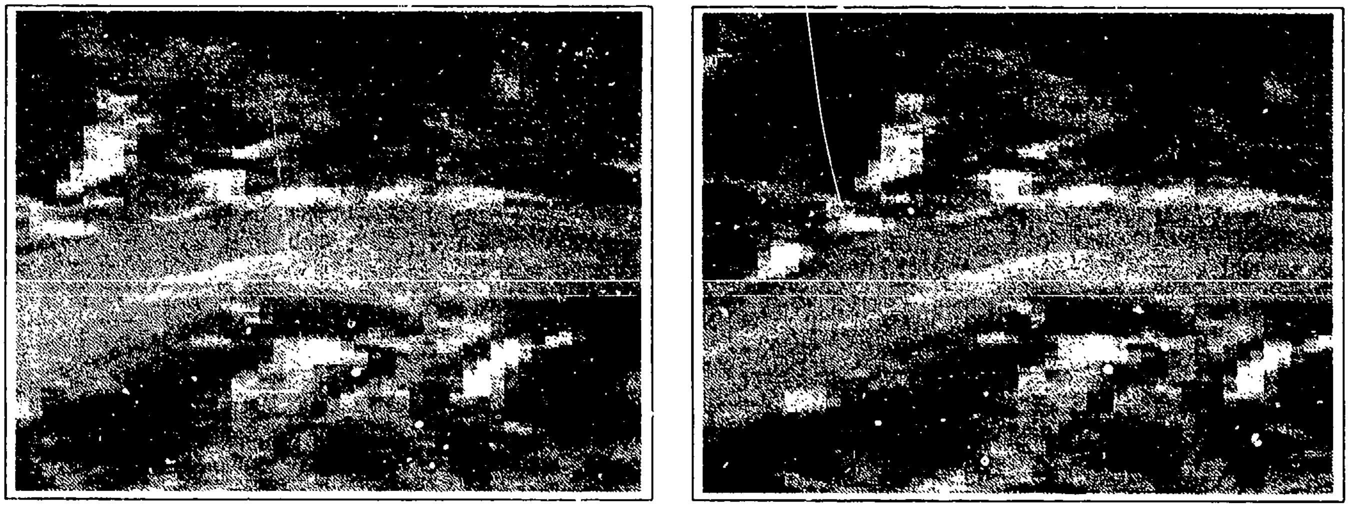
Noise and signal characteristics for this set of data are significantly different from those of the synthetically imaged CDC data. There is a great deal of small scale structure in the scene. The standard deviation of intensity variation was 25.603 here, with a standard deviation in first difference of $25.603\sqrt{2} = 38.0$. These measures account for the three levels of resolution reduction required to bring the noise down to an acceptable level. Figure 7-24 is a stereo plot of the intensity values of the left image of this pair (yet another interesting figure for the cross-eyed stereo freaks), with intensity being the $z$ component of the plot. It's startling just how much local intensity variation there is in these images. For comparison, Figure 7-25 shows a similar plot for the CDC data.



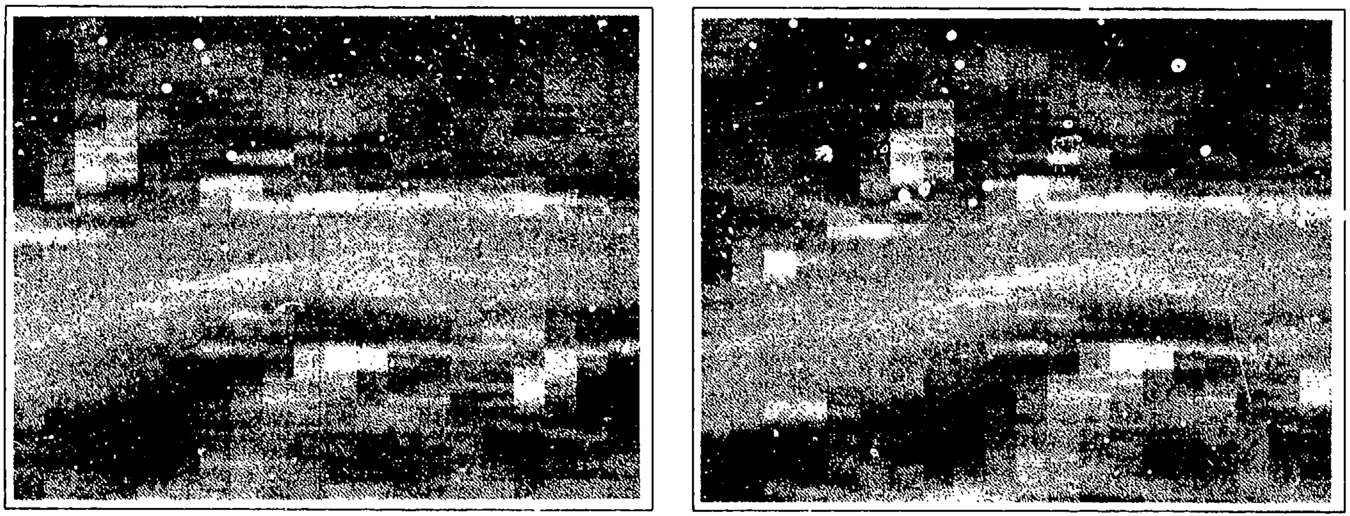
*NVL stereo pair of images — natural terrain* [168 × 200 × 9]
Figure 7-17



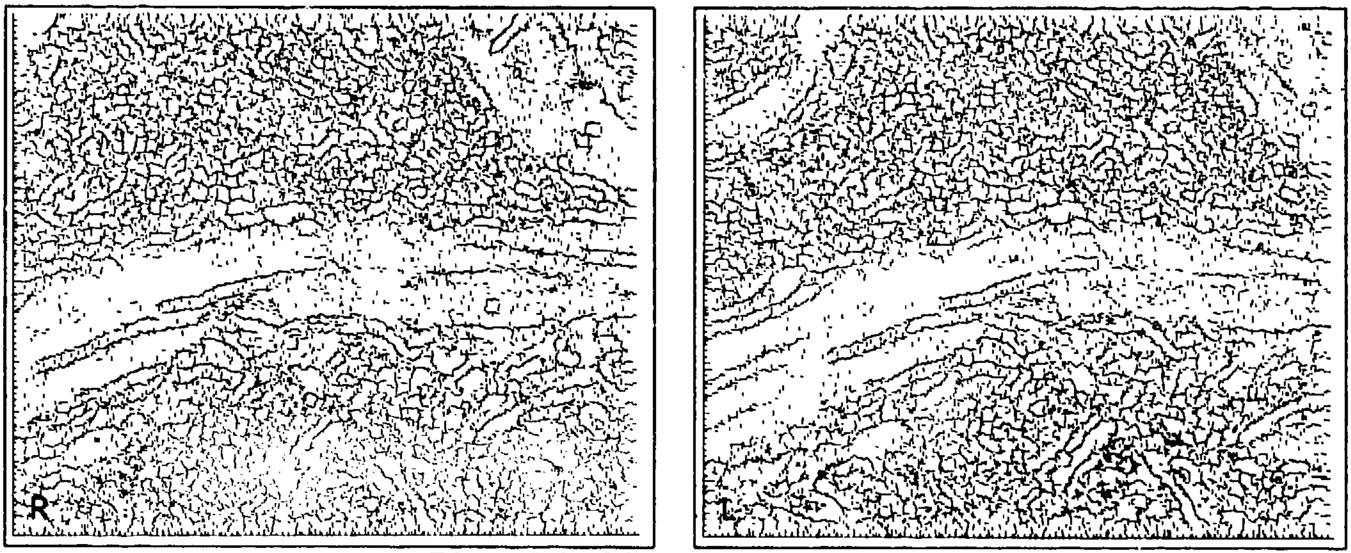
*First resolution reduction*
Figure 7-18



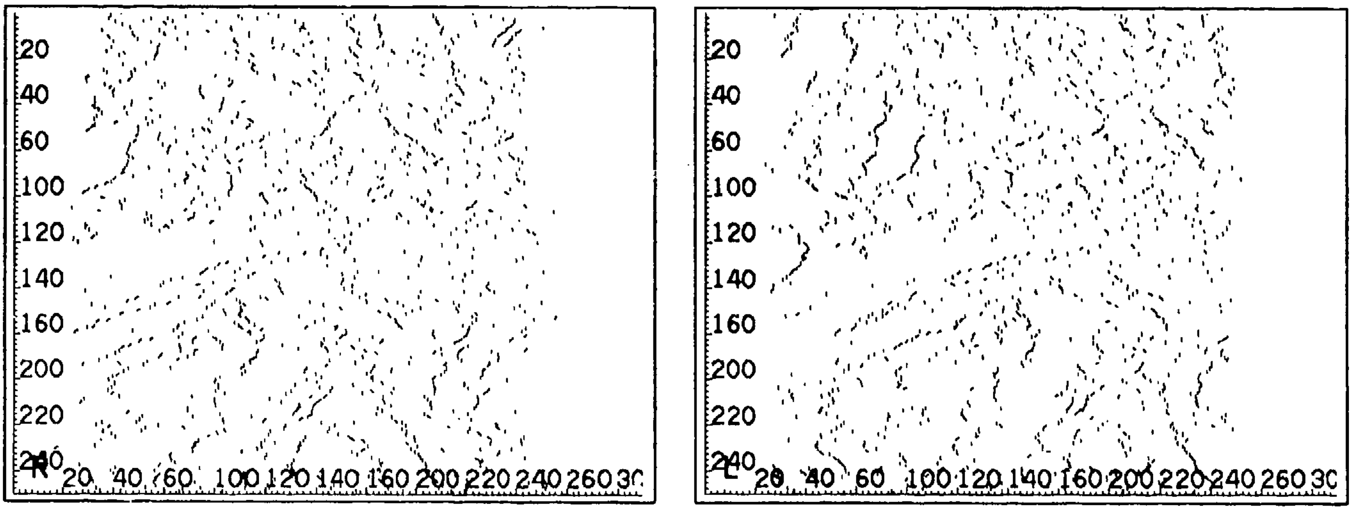
*Second resolution reduction*
Figure 7-19



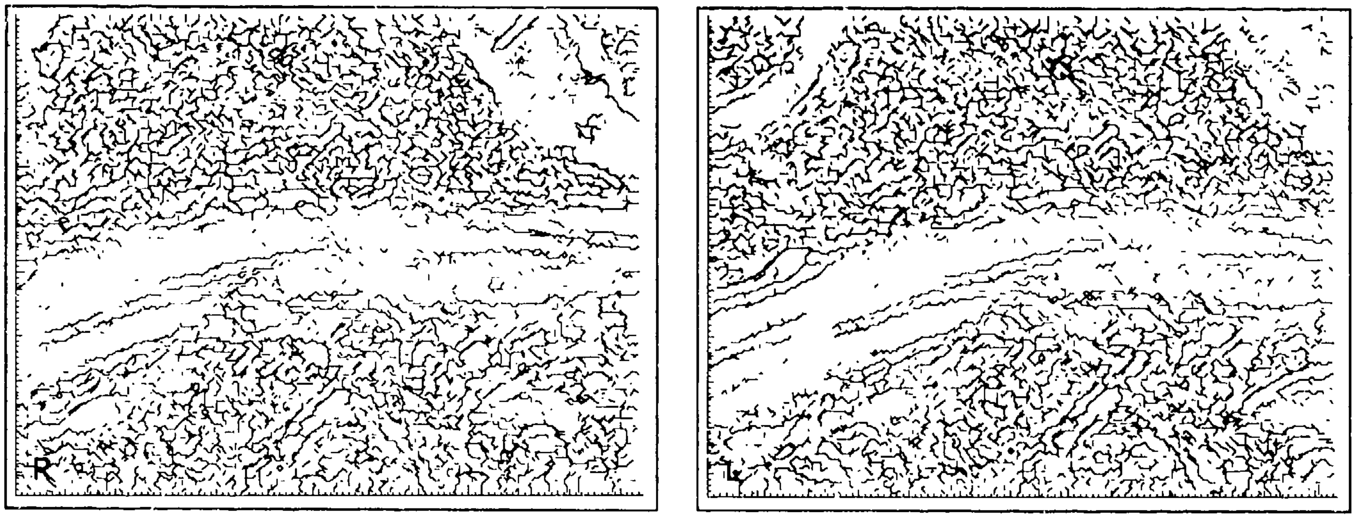
*Third resolution reduction*
Figure 7-20



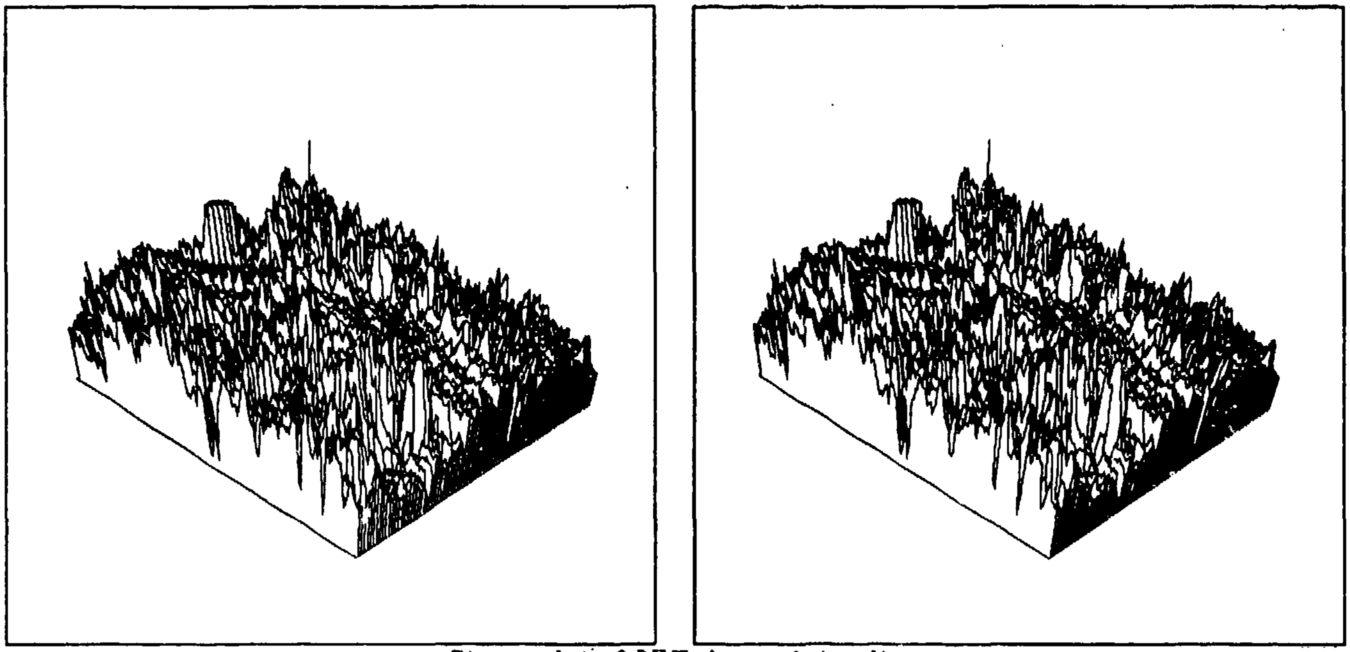
*Edges of the full resolution stereo pair*
Figure 7-21



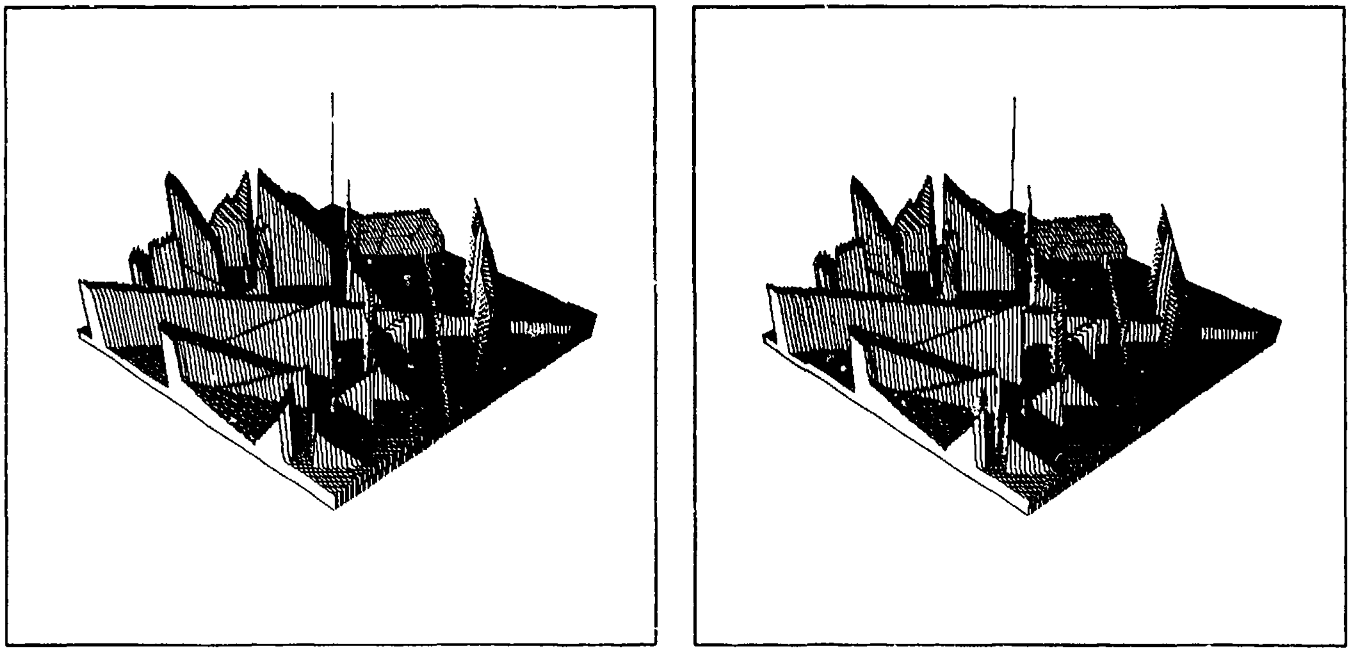
*Edges of the third resolution reduction*
Figure 7-22



*Edge connectivity at full resolution*
Figure 7-23

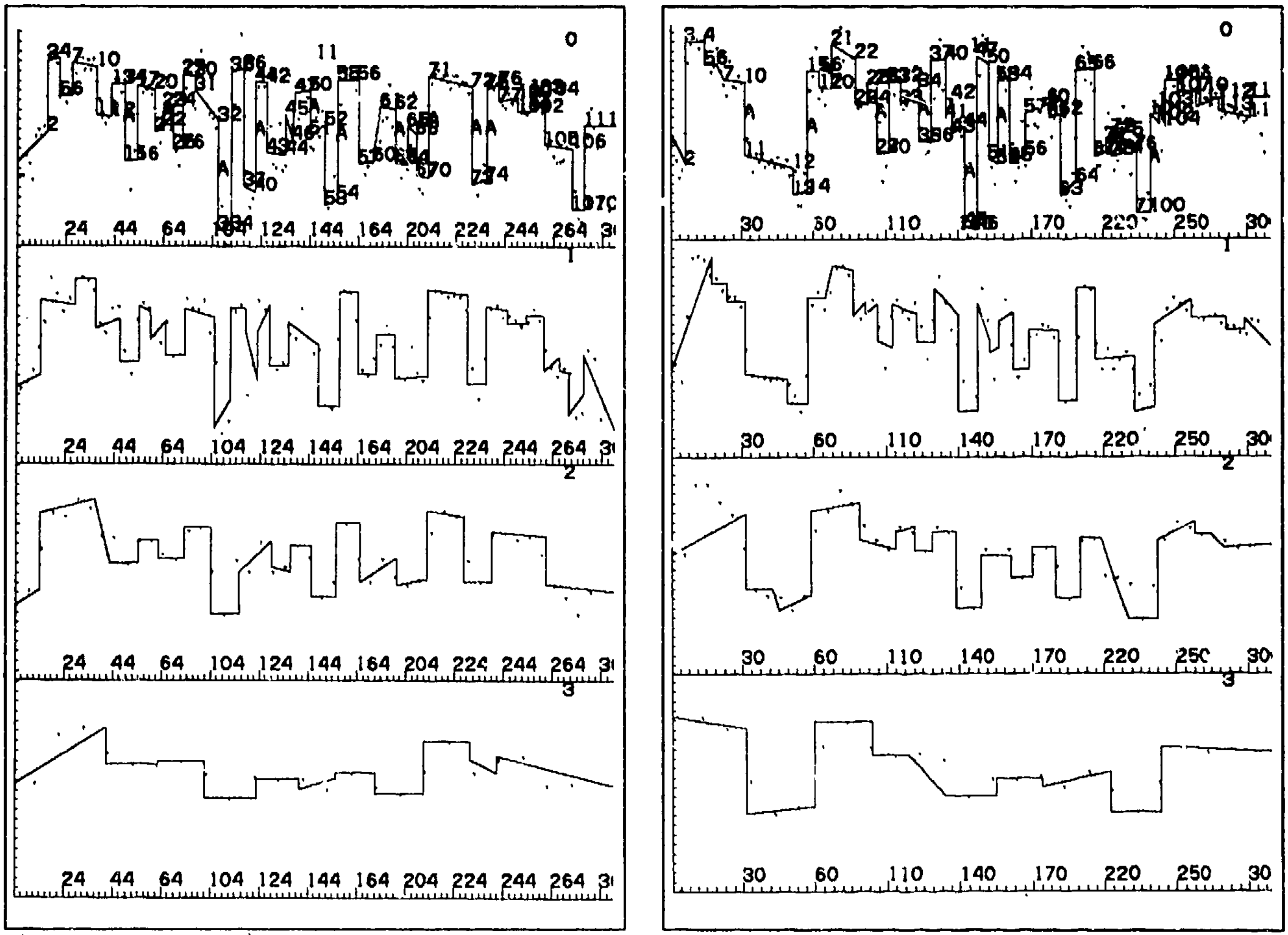


*Stereo plot of NVL image intensity*
Figure 7-24



*Stereo plot of CDC image intensity*
Figure 7-25

The next figure, Figure 7-26, shows the intensities and edges found along a single pair of corresponding lines in the successive resolution reductions of this imagery. Reduced resolution correspondences are found among edges in the bottom two figures, and these are then mapped up, through the intermediate resolution edges, to the full resolution edges at the top of the pair of figures. Full resolution correlation is then performed on the edges defined within corresponding intervals.



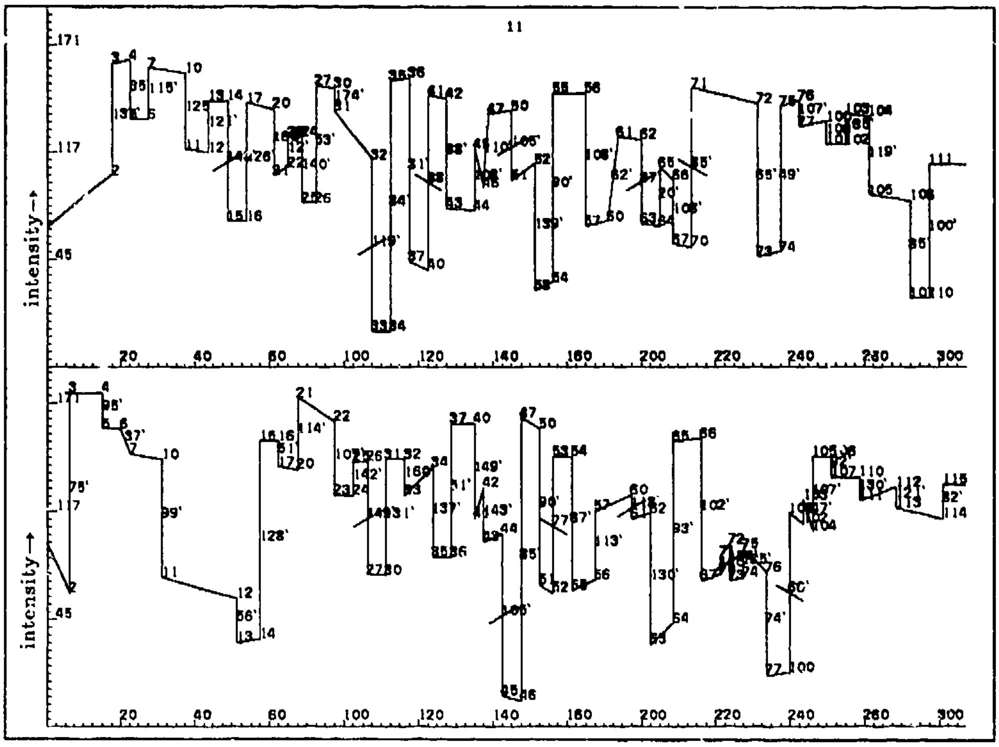
*Right and Left image line successive resolution reductions*
Figure 7-26

Figure 7-27 shows the top line of Figure 7-26 with half-edge indices marked. The reduced resolution correlation pairs edges:

{ (14, 26), (32, 44), (40, 52), (50, 60), (62, 72), (70, 100) }

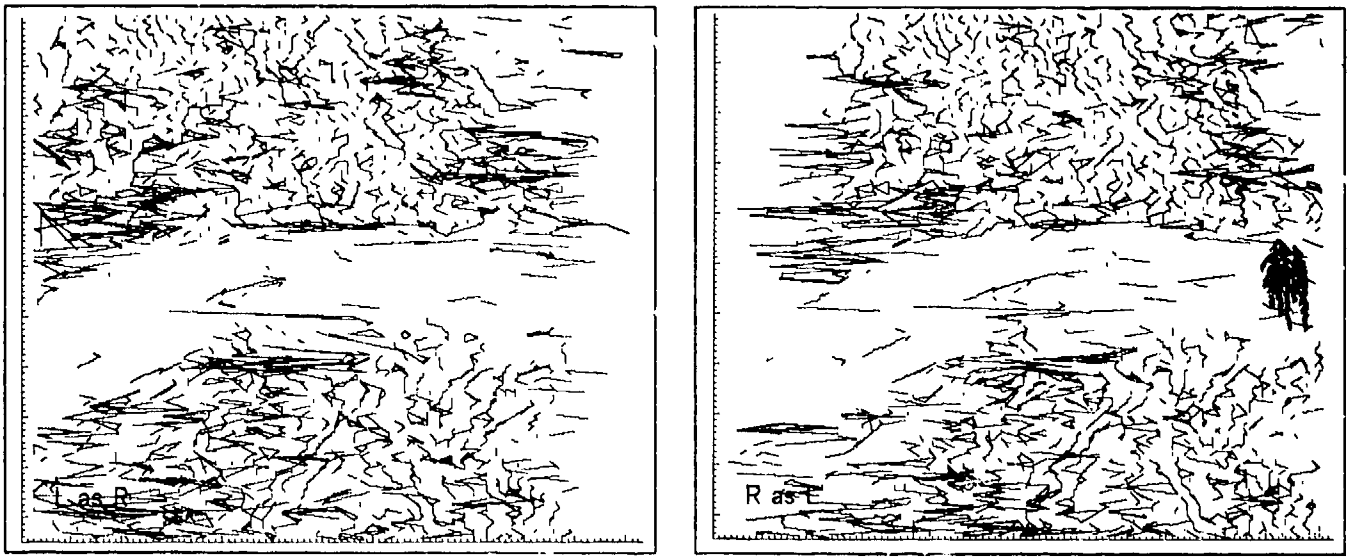
The full resolution correlation takes these correspondences, defining intervals for matching, and determines the pairings:

{ (3, 15), (4, 16), (5, 17), (6, 20), (7, 21), (10, 22), (11, 23), (12, 24),(13, 25), (14, 26),
(15, 27), (16, 30), (17, 31), (20, 32), (21, 33), (24, 34),(25, 35), (26, 36), (27, 37),
(30, 40), (31, 41), (32, 44), (33, 45), (34, 46),(35, 47), (36, 50), (37, 51), (40, 52),
(41, 53), (42, 54), (43, 55), (44, 56),(45, 57), (50, 60), (51, 61), (52, 62), (53, 63),
(54, 64), (55, 65), (56, 66),(57, 67), (66, 76), (67, 77), (70, 100), (71, 105),
(72, 106), (76, 110),(77, 111), (100, 112), (101, 113), (102, 114), (111, 115) }

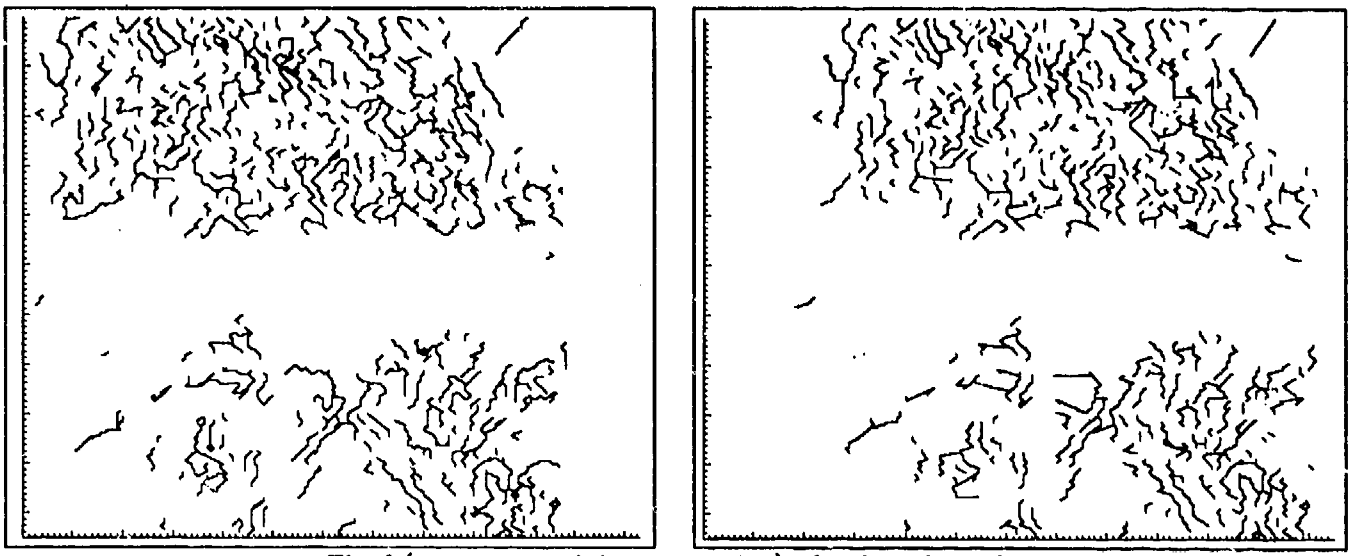


*Right and Left image line full resolution edges*
Figure 7-27

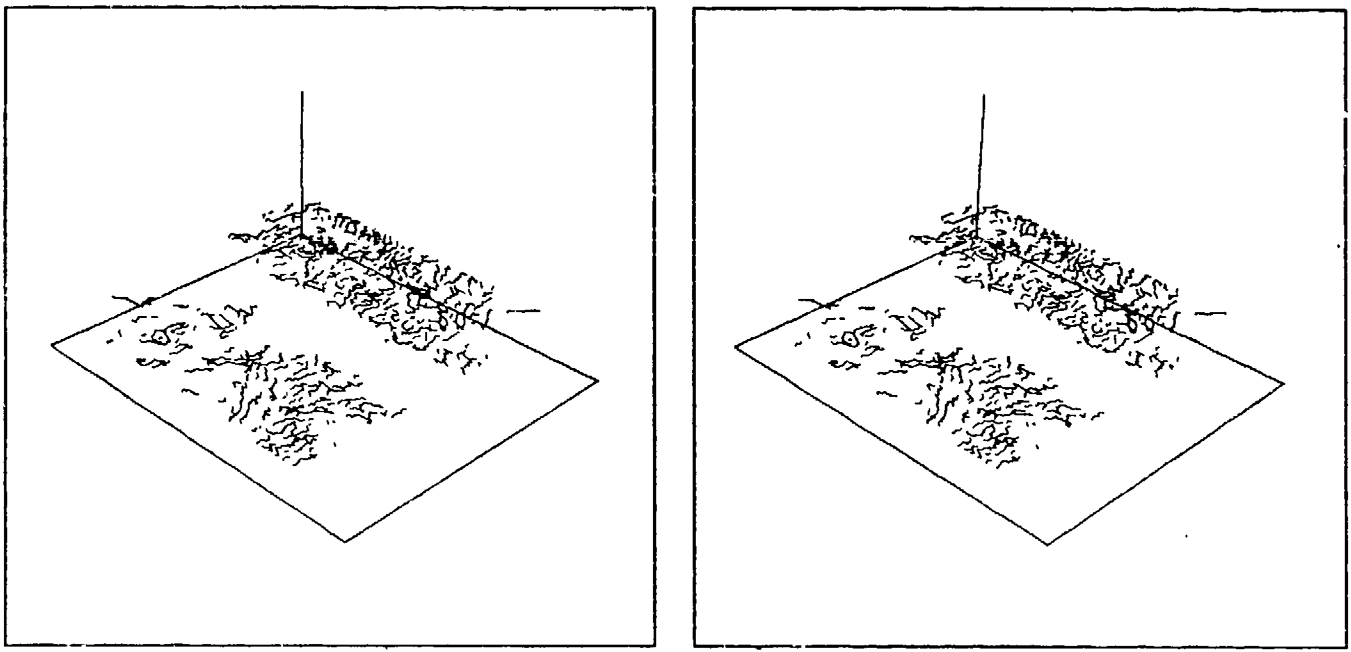
Preliminary results of the correlation are shown in Figure 7-28. Recall that horizontal lines jagging back and forth across the depiction are indicative of incorrect matches. Quite noticeably, there are many more errors in this correlation than there were for the comparable analysis of the CDC imagery. The hope is that the subsequent consistency enforcement process will be able to use image continuity to disambiguate the disparity jumps and produce a reliable set of edge matches. Figure 7-29 shows the results of the consistency enforcement process ··· a significant improvement. Figure 7-30 shows a 3-D perspective view of the connected edges (as was seen from directly overhead in Figure 7-29).



*Preliminary correlation results*
Figure 7-28

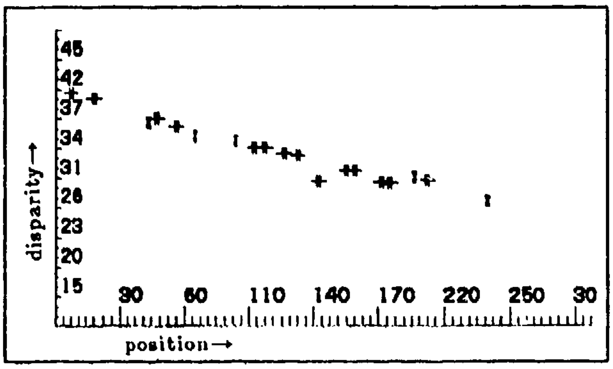


*Final (post-connectivity constraint) edge-based results*
**5700 half-edge correlate pairs**
Figure 7-29

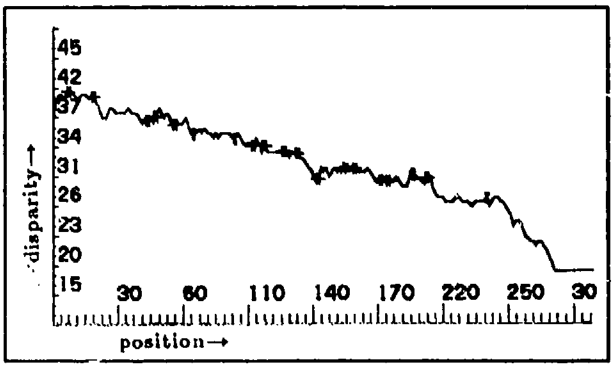


*Perspective view of connected edge elements*
Figure 7-30

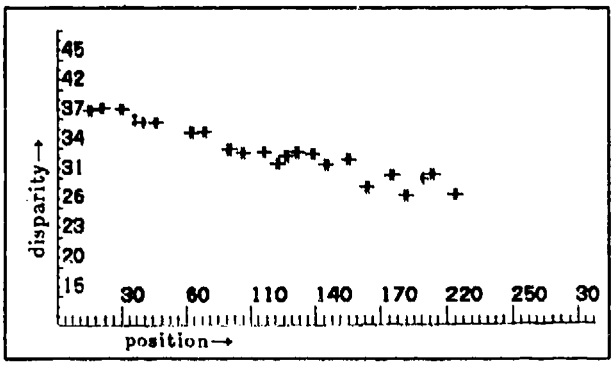
The correlation results at this stage form a ***template of constraints*** for the next stage of the processing, in which the interval-constrained edge-based and the interval-constrained intensity-based matchings attempt to complete the disparity array. Figures 7-31 and 7-33 show the correspondence of edges and disparities attained through these matchings for two sample image line-pairs. The depiction is identical to that of Figures 7-10 and 7-12, where the two types of edge mappings were indicated by the two different sorts of arrowheads. The intensity interpolation on these lines can be seen in Figures 7-32 and 7-34 (again, these displays are perspective, so verticals have varying horizontal components, but this isn't noticeable with the rolling nature of the terrain).



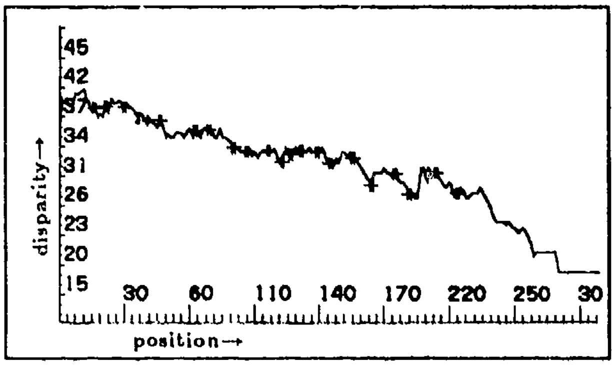
*Edge correspondences* (←→ *preliminary edge matches*; |< >|*subsequent edge matches*)
Figure 7-31



*Interpolated disparities*
Figure 7-32

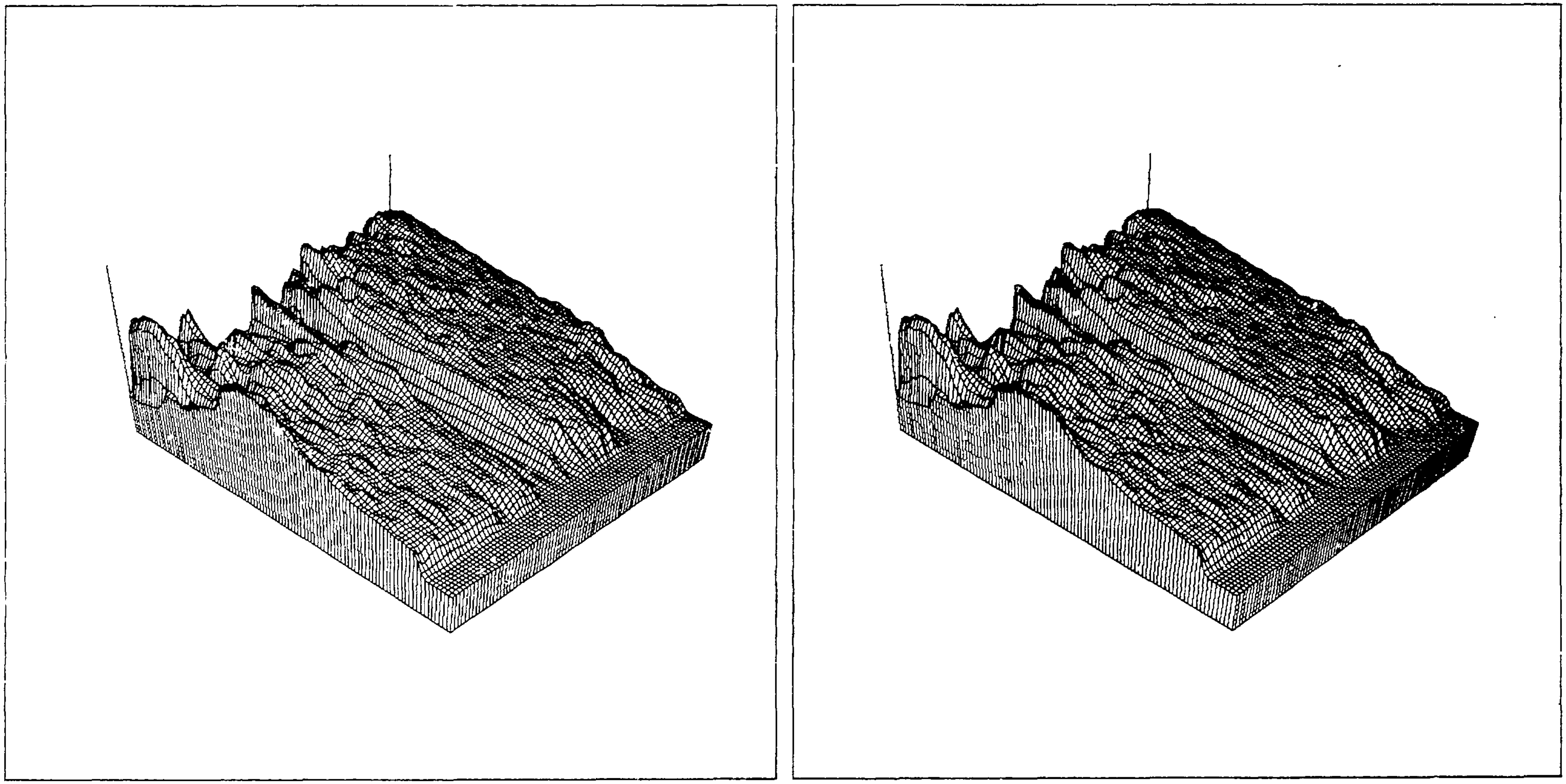


*Edge correspondences*
Figure 7-33

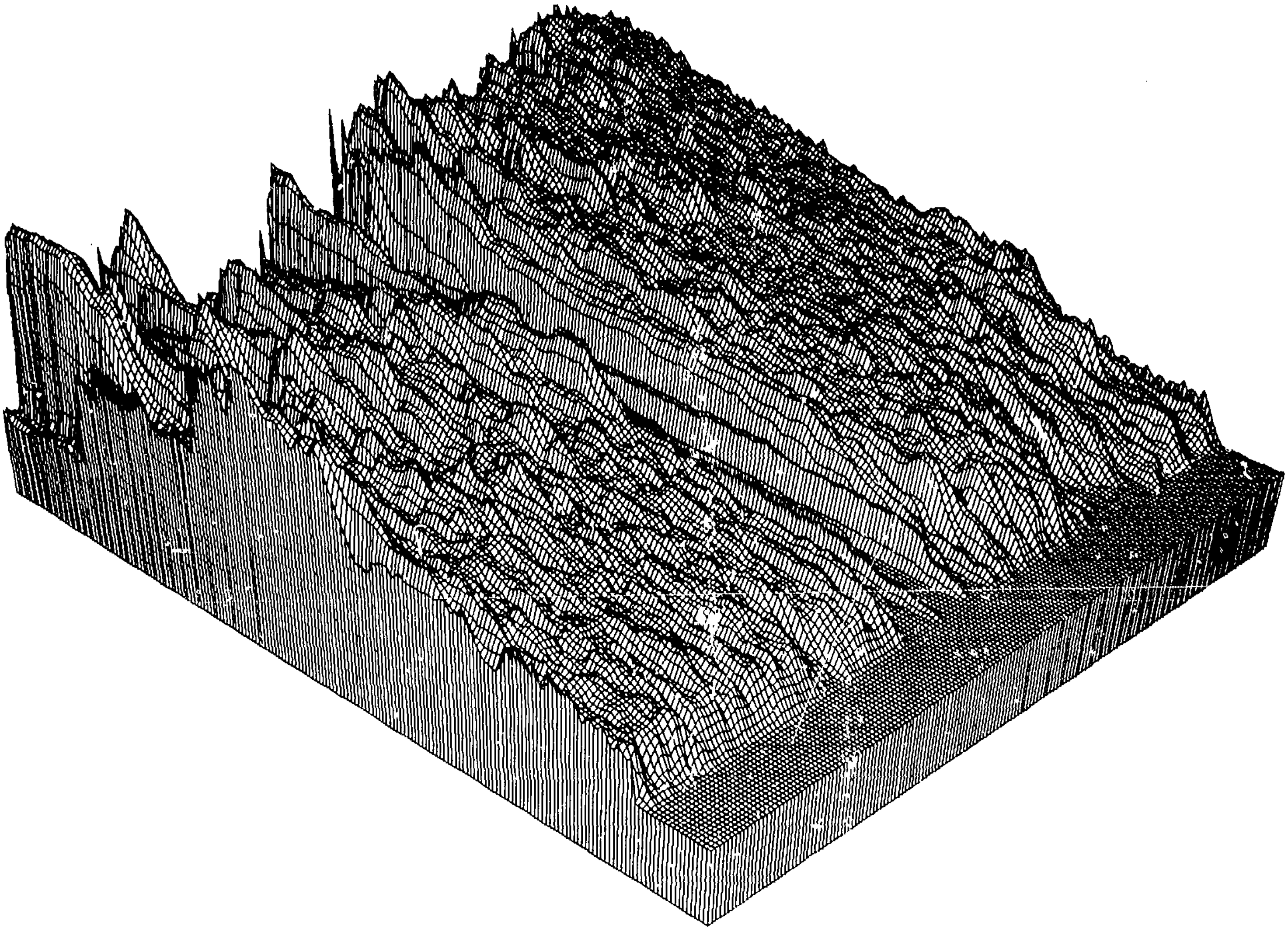


*Interpolated disparities*
Figure 7-34

The processing of this *interval-constrained edge*-based matching and the *interval-constrained intensity*-based matching result in the disparity map as shown at half resolution in Figure 7-35. Figure 7-36 shows a monocular view of the full resolution results of this processing.



*Perspective view of final edge and intensity correlation* — *NVL*
**after median filtering**
Figure 7-35



*Full resolution NVL plot (of Figure 7-35)*
Figure 7-36

# Chapter 8

# SUMMARY AND CONTRIBUTIONS TO THE FIELD

## 8.1 Further Considerations

Recall that the goals of this research were to develop a *robust, domain independent* stereo vision algorithm — one with a structure that would lend itself to a *parallel realization.*

[*Robustness*] The use of a *line-by-line coarse-to-fine* analysis capitalizing on the redundancy and broad frequency spectrum of grey-scale imagery and the accompanying *inter-line global* constraints provide for high noise-immunity, recovery from local correspondence errors, consistency at the level of global interpretation, and graceful degradation.

[*Domain Independence*] The examples shown in chapter 7 are from disparate domains. Having no monocular predispositions, beyond the dealing with edges, the system has nothing in it to bias the analysis toward a particular domain. Probabilistic measures used are those of *general* situations (although those for specific domains *could* be introduced if they were known and applicable). Testing on further imagery is expected to confirm the generality of the algorithm.

[*Parallel Implementable*] Estimates based on the run times of the two examples of chapter 7 suggest that the analysis proceeds at about 3 lines per second. It is thus expected that a parallel implementation on $n$ processors, fairly straightforward algorithmically from the current organization, would require something less than 0.5 seconds for an $n$-line by 256 element analysis with processors of the power of a DEC KL-10. A more likely early realization would be with something more modest, perhaps a successor of [Marks 1980], [Burr 1981], or [Lowry 1981].

More work is needed before this algorithm is ready for use in an integrated system. Primarily, more imagery data is needed in testing and demonstrating the comprehensiveness of the algorithm. The imagery shown in chapter 7 is a good beginning at indicating the power of the processing, but it can only suggest the potential — a broader and fuller image sampling is needed to be convincing of its generality.

Empirical analysis must also be made of the accuracy of the correlation algorithm. Digital terrain models (DTM's) with accompanying digital stereo imagery could provide the needed accuracy benchmarks for this. Unfortunately, acquiring DTM stereo imagery and databases has been problematic enough that I have not been able to include such an analysis in this report. Further work with this algorithm will certainly involve digital terrain model studies.

The set of parameters chosen for the various correlations should also be re-examined and perhaps augmented. Colour information may well be an extremely important addition to this. Although it has been shown through research with isoluminance that colour does not play a part in human primary stereopsis ([Gregory 1977]), there is no information-theoretic reason for so excluding it from a mechanized vision system. [Gregory 1977] points out that colour does function as a stimulus to *'contour'* stereo — the stereo from monocular cues, and my suspicion is that it will be a very powerful disambiguation metric for either correspondence process.

Refinement is needed in the spatial sampling used in both the resolution reduction and the lateral inhibition processes. This is of importance primarily for the resolution reductions, as the particular lateral inhibition operation implemented here is an artefact of the edge operator used which itself will surely be replaced by one with a better foundation ([Binford 1981]). Further two-dimensional analysis will also be needed in improving the ***constrained-interval intensity*** correlation. The errors seen in Figures 7-10 through 7-15 can be traced nearly without exception to the local ***line-by-line*** nature of its correlation. Much improvement with this is possible and expected were a more global analysis to be carried out.

## 8.2 Application of the Analysis

The research does not end with the development of an algorithm such as this. It is not a stand-alone process, but rather must serve as a provider of three-dimensional data for the modelling and recognition processes of a total machine vision system. Reference was made earlier to the importance of interfacing this sort of depth analysis to an object modelling system such as ACRONYM ([Brooks 1981b]). Reliable and accurate depth measurements would provide a new and invaluable capacity to the modelling system. Of course there are still many issues to be looked into for this. A few of the more obvious are:

- How is the depth map to be segmented for structure matching?
- What shape primitives are to be abstracted from the dense 3-space descriptions for object representation?
- Will the modelling scheme be able to direct the stereopsis process, suggesting monocular cues to guide the matching or providing cues to scene structure from the results of previous analyses?

Regardless of the path chosen for the implementation, the marriage of modelling and stereo analysis will come about — the benefits, if not mere necessity, of depth analysis makes this clear. A modelling system that can sense the world in 3-D can not only make better judgements about its environment, it can actively model that environment, forming solid descriptions of everything it encounters. The modelling will be able to do as we do — pick objects up, turn them about before its eyes, observe their static and dynamic characteristics, note similarities and differences with other objects seen and modelled before — doing all this on the basis of three-dimensional spatial structure. It is this generative aspect that makes the most exciting contribution to the modelling — objects will be modelled by being observed, with perhaps only the finest calibration measurements being added to the description manually. No longer would there be the necessity for object hand measurements and hand entry of object descriptors.

The automated stereopsis of this system will also bring advantage to terrain modelling and mapping. Its ability to handle both rolling terrain and the discontinuities of cultural site structures makes it applicable over a range of sensing situations not approached by current terrain mapping systems.

The most exciting aspects of vision research still lie ahead — at a sensing level, the incorporation of colour and monocular cueing and enhanced global analysis; at the level of segmentation, the recognition and clustering of surface shape primitives for coherent symbolic description; at the modelling level, the further extension or redesign of representational schemes to use this three-dimensional data; at the meta-modelling level, consideration of ways to describe shape and objects that will most effectively allow their recognition and manipulation.

# REFERENCES

[Arnold 1978] Arnold, R. David, "Local Context in Matching Edges for Stereo Vision," *Proceedings of the ARPA Image Understanding Workshop*, Boston, May 1978, 65–72. (*cited on pp.* 7,12,25)

[Arnold 1980] Arnold, R.D., and T.O. Binford, "Geometric Constraints in Stereo Vision," *Soc. Photo-Optical Instr. Engineers*, Vol. 238, Image Processing for Missile Guidance, 1980, 281–292. (*pp.* 38,40)

[Arnold 1982] Arnold, R. David, "Automated Stereo Perception," Department of Computer Science, Stanford University, forthcoming Ph.D. thesis, 1982. (*pp.* 60)

[Baker 1976] Baker, H. Harlyn, "Building Models of Three-Dimensional Objects," Master of Philosophy Thesis, The University of Edinburgh, Edinburgh, Scotland, 1976. (*pp.* 12)

[Baker 1980] Baker, H. Harlyn, "Edge Based Stereo Correlation," *Proc. ARPA Image Understanding Workshop*, University of Maryland, April 1980, 168–175. (*pp.* 25)

[Baker 1981a] Baker, H. Harlyn and Thomas O. Binford, "Depth from Edge and Intensity Based Stereo," *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, Vancouver, British Columbia, August 1981, 631–636. (*pp.* 21)

[Baker 1981b] Baker, H. Harlyn, "Depth from Edge and Intensity Based Stereo," University of Illinois, Ph.D. thesis, September 1981. (*pp.* ii)

[Baumgart 1974] Baumgart, Bruce G., "Geometric Modeling for Computer Vision," Ph.D. thesis, Stanford Artificial Intelligence Laboratory, AIM–249, October 1974. (*pp.* 12)

[Binford 1981] Binford, Thomas O., "Inferring Surfaces from Images," *Artificial Intelligence*, Vol. 17(1981), August 1981, 205–244. (*pp.* 19,33,86)

[Bishop 1975] Bishop, Peter O., "Binocular Vision," in *Adler's Physiology of the Eye*, 558–614, Robert A. Moses, editor, The C. V. Mosby Company, St. Louis, 1975. (*pp.* 1,2)

[Blakemore 1970] Blakemore, Colin, "A New Kind of Stereoscopic Vision," *Vision Research*, Vol. 10, 1970, 1181–1199. (*pp.* 50)

[Brooks 1981a] Brooks, Rodney A., ‘ Symbolic Reasoning Among 3-D Models and 2-D Images," Stanford Artificial Intelligence Laboratory, AIM–343, June 1981. (*pp.* v,20)

[Brooks 1981b] Brooks, Rodney A., "Symbolic Reasoning Among 3-D Models and 2-D Images," *Artificial Intelligence Journal*, Vol. 16, 1981. (*pp.* 86)

[Burr 1977] Burr, David J., "On computer stereo vision with wire frame models," Report R-805, University of Illinois, December 1977. (*pp.* 12)

[Burr 1981] Burr, D.J., Bryan Ackland, Neil Weste, "A High Speed Array Computer for Dynamic Time Warping," *IEEE International Conference on Acoustics, Speech and Signal Processing,* Atlanta, Ga., March 1981, 471–474. (*pp.* 85)

[Burt 1980] Burt, Peter and Bela Julesz, "A Disparity Gradient Limit for Binocular Fusion," *Science,* Vol. 208, No. 9, May 1980, 615–617. (*pp.* 17,29)

[Degryse 1980] DeGryse, Donald G. and Dale J. Panton, "Syntactic Approach to Geometric Surface Shell Determination," *Soc. Photo-Optical Instrumentation Engineers,* Vol. 238, August 1980, 264–272. (*pp.* 12,13,15,16,17)

[Forney 1973] Forney, G. David Jr., "The Viterbi Algorithm," *Proc. IEEE,* Vol. 61, No. 3, March 1973, 268–278. (*pp.* 13,45)

[Friedman 1980] Friedman, S.J., editor, Manual of Photogrammetry, American Society of *Photogrammetry,* C. C. Slama, editor-in-chief, 1980. (*pp.* 8)

[Frisby 1977] Frisby, John P. and John E.W. Mayhew, "Global Processes in Stereopsis: Some Comments on Ramachandran and Nelson(1976)," *Perception,* Vol. 6, 1977, 195–206. (*pp.* 15)

[Gennery 1980] Gennery, Donald B., "Modelling the Environment of an Exploring Vehicle by Means of Stereo Vision," Ph.D. thesis, Stanford Artificial Intelligence Laboratory, AIM–339, June 1980. (*pp.* v,6,7,11,12,17,18,19)

[Gibson 1950] Gibson, James J., *"The Perception of the Visual World,"* The Riverside Press, Houghton Mifflin Co., 1950. (*pp.* 4)

[Gimel'farb 1972] Gimel'farb, G.L., V.B. Marchenko, and V.I. Rybak, "An Algorithm for Automatic Identification of Identical Sections on Stereopair Photographs," *Kybernetica* (translations) No. 2, March–April 1972, 311–322. (*pp.* 8,13)

[Gregory 1977] Gregory, Richard L., "Vision with Isoluminant Colour Contrast: 1. A Projection Technique and Observations," *Perception,* Vol. 6, 1977, 113–119. (*pp.* 4,85)

[Grimson 1980] Grimson, William Eric Leifur, "From Images to Surfaces: A Computational Study of the Human Early Visual System," William Eric Leifur Grimson, MIT Press, 1981. An earlier version was printed as: "Computing Shape Using a Theory of Human Stereo Vision," Department of Mathematics, MIT, Ph.D. thesis, June 1980. (*pp.* 12,14,17,18,20)

[Hallert 1960] Hallert, Bertil, *"Photogrammetry, Basic Principles and General Survey,"* McGraw-Hill Book Company Inc., 1960. (*pp.* 20)

[Hannah 1974] Hannah, Marsha Jo, "Computer Matching of Areas in Stereo Images," Ph.D. thesis, Stanford Artificial Intelligence Laboratory, AIM–239, July 1974. (*pp.* 6,7,9,21)

[Henderson 1979] Henderson, Robert L., Walter J. Miller, C.B. Grosch, "Automatic Stereo Recognition of Man-Made Targets," *Society of Photo-Optical Instrumentation Engineers,* Vol. 186, *Digital Processing of Aerial Images,* August 1979. A more complete description is available as: "Geometric Reference Preparation Interim Report Two: The Broken Segment Matcher," Henderson, R.L., Rome

Air Development Centre, Rome, New York, RADC–TR–79–80, April 1979. (*pp.* v,6,12,15,16,17,19,45)

[Hueckel 1971] Hueckel, M., "An Operator which Locates Edges in Digital Pictures," *Journal of the ACM*, Vol. 18, no.1, January 1971, 113–125. Erratum in 21, 350, 1974. (*pp.* 12)

[Julesz 1971] Julesz, Bela, *"Foundations of Cyclopean Perception,"* Chicago, University of Chicago Press, 1971. (*pp.* 3,4)

[Julesz 1976] Julesz, Bela, "Global Stereopsis: Cooperative Phenomena in Stereoscopic Depth Perception," in *Handbook of Sensory Physiology VIII*, R. Held, H. Leibovitz and H-L. Teuber, editors, Springer, Berlin, 1976. (*pp.* 1,15,47)

[Kelley 1970] Kelley, Michael D., "Visual Recognition of People by Computer," *Stanford Artificial Intelligence Laboratory, AIM-130, CS-168,* Ph.D. thesis, 1970. (*pp.* 19,20)

[Kelly 1977] Kelly, R.E., P.R.H. McConnell, and S.J. Mildenberger, "The Gestalt Photomapping System," *Journal of Photogrammetric Engineering and Remote Sensing,* Vol. 43, 1407, 1977. (*pp.* v,8)

[Levine 1973] Levine, Martin D., Douglas A. O'Handley, Gary M. Yagi, "Computer Determination of Depth Maps," *Computer Graphics and Image Processing,* 2, 1973, 131–150. (*pp.* 6,7,8,17,18,19)

[Liebes 1981] Liebes Jr., S., "Geometric Constraints for Interpreting Images of Common Structural Elements: Orthogonal Trihedral Vertices," *Proceedings of the ARPA Image Understanding Workshop,* 1981. (*pp.* 16)

[Lowry 1981] Lowry, Michael R. and Allan Miller, "A General Purpose VLSI Chip for Computer Vision with Fault-Tolerant Hardware," *Proc. ARPA Image Understanding Workshop,* Washington D. C., April 1981, 184–187. (*pp.* 85)

[Marks 1980] Marks, Philip, "Low-Level Vision Using an Array Processor," *Computer Graphics and Image Processing,* No. 14, 1980, 281–292. (*pp.* 85)

[Marr 1976] Marr, D. and T. Poggio, " Cooperative Computation of Stereo Disparity," *Science,* Vol. 194, October 1976, 283–287. (*pp.* 3,25)

[Marr 1977] Marr, D. and T. Poggio, "A Theory of Human Stereo Vision," MIT Artificial Intelligence Memo No. 451, November 1977. (*pp.* 3,12,14,20)

[Marr 1979] Marr, D., T. Poggio, S. Ullman, "Bandpass Channels, Zero-crossings, and Early Visual Information Processing," *Journal of the Optical Society of America.* 1979. (*pp.* 4)

[Mayhew 1981] Mayhew, John E.W. and John P. Frisby, "Computational and Psychological Studies Towards a Theory of Human Stereopsis," *Artificial Intelligence Journal,* Vol. 16, 1981. (*pp.* 3,15,25,54)

[Moore 1979] Moore, Roger K., "A Dynamic Programming Algorithm for the Distance Between Two Finite Areas," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 1, No. 1, January 1979, 86–88. (*pp.* 60)

[Moravec 1980] Moravec, Hans P., "Obstacle Avoidance and Navigation in the Real World by a Seeing Robot Rover," Stanford Artificial Intelligence Laboratory, AIM-340, Ph.D. thesis, September 1980. (*pp.* v,6,10,11,12,18,20)

[Mori 1973] Ken-Ichi Mori, Masatsugu Kidode, Haruo Asada, "An Iterative Prediction and Correction Method for Automatic Stereocomparison," *Computer Graphics and Image Processing*, 2, 1973, 393-401. (*pp.* 7,9)

[Nishihara 1981] Nishihara, H.K., and N.G. Larson, "Towards a Real Time Implementation of the Marr and Poggio Stereo Matcher," *Proceedings of the ARPA Image Understanding Workshop*, May 1981, 114-120. (*pp.* 12,19)

[Panton 1978] Panton, Dale J., "A Flexible Approach to Digital Stereo Mapping," *Photogrammetric Engineering and Remote Sensing*, **Vol.** 44, **No.** 12, December 1978, 1499-1512. (*pp.* v,7,10,17,18,19,21)

[Panton 1981] Panton,D.L., C.B. Grosch, D.G. DeGryse, J. Ozils, A.E. LaBonte, S.B. Kaufmann, L. Kirvida, "Geometric Reference Studies," RADC-TR-81-182, Final Technical Report, July 1981. (*pp.* 12,13,15,16,17)

[Richards 1970] Richards, Whitman, "Stereopsis and Stereoblindness," *Exp. Brain Research*, **Vol.** 10, 1970, 380-388. (*pp.* 1)

[Rubin 1980] Rubin, Steven M., "Natural Scene Recognition Using Locus Search," *Computer Graphics and Image Processing*, **Vol.** 13, **No.** 4, August 1980, 298-333. (*pp.* 46)

[Ryan 1979] Ryan, T.W., R.T. Gray, and B.R. Hunt, "Prediction of Correlation Errors in Stereo-Pair Images," SIE/DIAL-79-002. (*pp.* 6)

[Ryan 1980] Ryan, Thomas W., and B.R. Hunt, "The Prediction of Accuracy in Digital Cross-Correlation of Stereo-Pair Images," *Soc. Photo-Optical Instr. Engineers*, **Vol.** 219, Electro-Optical Technology for Autonomous Vehicles, 1980. (*pp.* 6)

[Saye 1975] Saye, Ann and John P. Frisby, "The Role of Monocularly Conspicuous Features in Facilitating Stereopsis from Random-Dot Stereograms," *Perception*, **Vol.** 4, 1975, 159-171. (*pp.* 4,15)

[Scarano 1976] Scarano, Frank A., "A Digital Elevation Data Collection System," *Photogrammetric Engineering and Remote Sensing*, **Vol.** 42, **No.** 4, 489, April 1976. (*pp.* 8)

[Schumer 1979] Schumer, Robert A., "Mechanisms in Human Stereopsis," Ph.D. thesis, Department of Psychology, Stanford University, 1979. (*pp.* 4,18,52)

[Walk 1961] Walk, Richard D. and Eleanor J. Gibson, "A Comparative and Analytic Study of Visual Depth Perception," *Psychological Monographs*, **Vol.** 75, **No.** 15, 1961, 2-34. (*pp.* 5)

[Wilson 1978a] Wilson, Hugh R., "Quantitative Prediction of Line Spread Function Measurements: Implications for Channel Bandwidths," *Vision Research*, **Vol.** 18, 1978, 493-496. (*pp.* 3,4,20)

[Wilson 1978b] Wilson, Hugh R. and James R. Bergen, "A Four Mechanism Model for Threshold Spatial Vision," University of Chicago, April 1978. (*pp.* 3)

[Yonas 1978] Yonas, Albert, Wallace T. Cleaves, and Linda Pettersen, "Development of Sensitivity to Pictorial Depth," *Science*, **Vol.** 200, April 7, 1978, 77-79. (*pp.* 5)